Knowledge and Information Systems (2021) 63:143–173
https://doi.org/10.1007/s10115-020-01514-8

REGULAR PAPER

Multiscale Laplacian graph kernel combined with lexico-syntactic patterns for biomedical event extraction from literature

Sabenabanu Abdulkadhar[1] · Balu Bhasuran[2] · Jeyakumar Natarajan[1,2]

Received: 18 May 2019 / Revised: 22 September 2020 / Accepted: 27 September 2020 /
Published online: 24 October 2020

Abstract

Bio-event extraction is an extensive research area in the field of biomedical text mining, this focuses on elaborating relationships between biomolecules and can provide various aspects of their nature. Bio-event extraction plays a vital role in biomedical literature mining applications such as biological network construction, pathway curation, and drug repurposing. Extracting biological events automatically is a difficult task because of the uncertainty and assortment of natural language processing such as negations and speculations, which provides further room for the development of feasible methodologies. This paper presents a hybrid approach that integrates an ensemble-learning framework by combining a Multiscale Laplacian Graph kernel and a feature-based linear kernel, using a pattern-matching engine to identify biomedical events with arguments. This graph-based kernel not only captures the topological relationships between the individual event nodes but also identifies the associations among the subgraphs for complex events. In addition, the lexico-syntactic patterns were used to automatically discover the semantic role of each word in the sentence. For performance evaluation, we used the gold standard corpora, namely BioNLP-ST (2009, 2011, and 2013) and GENIA-MK. Experimental results show that our approach achieved better performance than other state-of-the-art systems.

Keywords Bio-event extraction · Graph kernel · Multiscale Laplacian Graph kernel · Pattern matching rule engine · BioNLP-ST · GENIA-MK

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10115-020-01514-8) contains supplementary material, which is available to authorized users.

✉ Jeyakumar Natarajan
n.jeyakumar@yahoo.co.in

[1] Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, TamilNadu 641046, India

[2] DRDO-BU Center for Life Sciences, Bharathiar University Campus, Coimbatore, TamilNadu 641046, India

1 Introduction

Advances in both biological and computational methods act as the catalyst for a large number of publications, especially in the biomedical domain [1]. Life science research outputs are widely disseminated as scientific articles, which can act as a source for knowledge discovery [2]. Recently biomedical text mining applications are developed using this literature with a focus on biological and clinical domain areas such as screening of clinical trials, pharmacogenetics, reaction detection and repurposing of drugs [3].

Initial efforts on text mining in the biomedical domain had a major focus on fundamental tasks like categorizing bio-entities (genes, proteins, diseases, and drugs) and extracting binary relationships (protein–protein interaction, gene–disease associations and disease–disease associations) between the entities [4]. Extracting relations from biomedical literature is a significant task in the area of semantic mining of text [5]. Some of the recent relation extraction strategies applied to various biomedical problems such as protein–protein interactions (PPIs) [6], gene–disease associations [7], chemical-induced disease (CID) [8] and chemical–disease relation (CDR) [9]. Biomedical relation classification task focusing on PPI and drug–drug interaction [10] shows the importance and applications of relation extraction from the literature.

Following the success of the relation extraction task, the next focus is on to extract related biomolecular events from the text. In general, bio-event is the textual event specialized for the biomedical domain and dynamic bio-relation involving one or more participants, and these participants can be bio-entities or bio-events and are usually each assigned a semantic role like the theme and cause [11, 12]. Bio-event extraction can help us to understand certain biological processes such as pathway reconstruction [13], semantic search [14], association mining for knowledge discovery, and bioprocess extraction [15]. Automatically, extracting events from the biomedical text is a challenging task because of the uncertainty and assortment of NLP processing such as negations and speculations, which occur in the biological text and can lead to misunderstanding and incorrect interpretation [11, 12].

The bio-event extraction process consists of two common steps, trigger detection and argument detection. Identifying trigger words comprises the detection of event triggers and their types, as quantified by the selected ontology [11]. Argument detection, known as edge detection or event theme construction is the process of detecting arguments for the events. The arguments can be named entities (genes, proteins, diseases) or events represented by trigger words [11, 12, 16]. Consider the following example below.

Example: PMCID: 1310901

Original Sentence: Down-regulation of interferon regulatory factor 4 gene expression in leukemic cells.

Tagged Sentence: <trigger>***Down-regulation*** </trigger> of <theme>***interferon regulatory factor 4*** </theme><trigger>***gene expression*** </trigger> in leukemic cells.

Here the trigger words '*downregulation'* and '*expression*' denote the two events - regulation and gene expression, and the gene '*interferon regulatory factor 4*' is the theme representing the argument in the sentence.

There has been a wider acceptance of the notion that biomolecular events can play a crucial role in molecular mechanisms of diseases and can be linked with interactions in pathways and networks [12, 16]. Due to this and other various reasons, notable shared task community challenges BioNLP-ST (Biomedical Natural Language Processing Shared Task) in 2009 [16], 2011 [17], 2013 [18] and 2016 [19] were organized specifically in focus on biomolecular event extraction from the literature. The core problem in these tasks was the extraction of biomolecular events from standard datasets, which is based on the GENIA corpus [20]. The GENIA corpus enriched with domain-specific meta-knowledge and it was named as GENIA-MK (Meta-knowledge) corpus [21]. The GENIA-MK corpus contains human curated annotations of 9,372 sentences from 1000 abstracts in which 36,858 typed, complex and nested events were represented [21]. Recently Zerva et al. [22] proposed a hybrid approach combining a random forest with generic rule patterns, which uses dependency between trigger words and cues of the uncertainty events and achieved an F-Score of 88% in the GENIA-MK corpus.

1.1 Background

Different text-mining approaches have been developed utilizing techniques such as rule-based [23], dictionary based [24], machine learning [25], and hybrid approaches [26]. In particular, the Support Vector Machines algorithms with rule-based or dictionary-based approaches are widely used in extracting biomolecular events [27]. In spite of several existing approaches, the challenge is still open and leaves space for improvement. For example, pattern matching and dictionary-based approaches achieved moderate results in complex event extraction processes such as regulation, negative regulation, and positive regulation [11]. Machine learning based studies [25] employed different strategies such as kernel-based learning [28, 29], deep learning based [30–32], graph-based learning [33–41] and hybrid approaches [26] to extract the biomedical events efficiently.

Recently, the enriched graph-based features played an important role to extract the events from the text and created the best systems for the classification of biological events [42]. The advantage of using graph-based approaches for event extraction includes the use of structural properties of the sentence such as semantic and syntactic features, path features, and similarity features. This was briefly explained in the review [42]. Earlier, various graph-based approaches like subgraph mining [43–45], random walk [46], shortest paths [47], subgraph matching [39–41, 48] and hybrid methods [49] were introduced to extract the biomedical events from the literature.

Subgraph mining is the process of extracting the important concepts from the graph [43–45]. Random walk explains the path consists of random steps between one node (bio-entity) to another node (bio-event) in the graph [46]. The shortest path is the shortest optimized path between two nodes (*entity* and *event*) [47]. The graph matching techniques are utilized to find whether one text could be inferred from another by using the dependency parsing of the two texts [39]. Subgraph matching techniques are utilized to extract the maximum common subgraph between two graphs [39–41]. On the other side, kernel-based approaches integrated with graphs produced efficient results in relation extraction tasks [50, 51]. A graph kernel was generated using dependency parsing techniques in which each graph contains the dependency structure and the linear order of the words [52]. In this study, we employed a special graph kernel named Multiscale Laplacian Graph (MLG) kernel [53] integrated with the linear feature-based kernel to extract the biological events from the text. The MLG-Kernel was used to compare the structure of the graph at multiple different scales. The motivation behind employing MLG is that it not only captures the topological relationships between the individual event nodes but also identifies the topological relationships between the subgraphs [53]. The following section briefly describes state-of-the-art approaches for the task of biomedical event extraction.

1.2 Related work

Bjorne et al. [33] used n-gram features and shortest path syntactic dependencies between event arguments and rule-based graph pruning to extract the events and attained the F-score 51.95% in the BioNLP-ST-2009 task dataset. The disadvantage of this approach is the lowest trigger detection performance on the test set. In 2013, BioNLP-ST, Bjorne and Salakoski [34] presented an automated event extraction system named TEES 2.1. It is a machine learning based tool for extracting text bound graphs from natural language articles, they represent both binary relations and events with a unified graph format where named entities and triggers are nodes and relations and event arguments are edges and reported an F-score of 50.74%. The lack of using learning rules caused defects in the argument detection phase; for example, consider an event with multiple optional arguments, such as *Cell differentiation* from the CG task with 0–1 AtLoc argument and 0–1 Theme arguments. While it can be possible that such an event can exist without any arguments at all, it is often the case that at least one of the optional arguments must be present. Hakala et al. [35] used graph represented features including paths connecting nested events and the occurrence of a pair of entities such as gene, protein in general subgraphs mined from external PubMed and PMC abstracts reported the best F-score of 50.97% in BioNLP-ST-2013. The main limitation of this system is that it increases only precision not recall.

In BioNLP-ST-2011 Riedel et al. [36], extracted event arguments by scoring candidate subgraphs to rank event pairs and achieved the F-score of 57.46%. In this system, they employed stacking and the UMass model (trained model which consists of trigger labels, events arguments and protein pairs) to extract the events. Stacking led to better performance in this system but a combination of stacking with the UMass model caused slight variation in the performance on the test sets. McClosky [54] converted annotated event structure in the training data to an event dependency graph that takes entities (event arguments) as vertices and edges and attained the F-score of 50% in BioNLP-ST-2011. Riedel and McCallum [55] implemented stacking procedure and combined their approach with McClosky [54] extracted event arguments by scoring candidate subgraphs to rank event arguments and achieved the F-score of 56.05% in the BioNLP-ST-2011 dataset; the limitation of this approach is that it is harder to extract full text events.

Liu et al. [39, 40] implemented Exact Subgraph Matching and Approximate Subgraph Matching (ESM/ASM) approaches to extract the events from the literature efficiently. In their method, they applied ESM/ASM from sentence graphs to event graphs, employed a distance metric to every vertex of the subgraphs, and attained the F-score of 51.12% in the BioNLP-ST-2011 dataset. The lack of post-processing rules and inconsistencies in the gold annotation caused more false positives and false negatives in this system. Liu et al. [41] further improved their ESM/ASM based approach with the distributional similarity model (DSM), optimized graph features, and attained the F-score of 55.09% in BioNLP-ST- 2013. The limitation of this approach is low recall due to 'Site' entity recognition.

Apart from the above graph-based approaches, recently different classification approaches were also deployed to extract the biomedical events efficiently [30, 56–58]. Some of the notable works are discussed here. Munkhdalai et al. [56] proposed a new semi-supervised learning method which was named self-training in significance space (STSS) to solve the imbalanced data problem and attained the F-score of 54.30% in BioNLP-ST-2011.The system performance is lower in terms of F-measure because of the computational requirements. Wang et al. [30] presented a multiple distributed representation method which combines dependent context formed by word embedding with task-based features from biomedical text and fed it to deep learning models and achieved the F-scores 59.94%, 55.20%, and 50.12%

in BioNLP-ST-2009, 2011, 2013 datasets, respectively; this method still needs manually designed features, which limits the power of generalization. Li et al. [57] used an optimization method named dual decomposition method along with dependency parse based rich features, unsupervised word features and extracted the events with F-scores 56.09% and 53.19% in BioNLP-ST- 2009, 2013. Recently, Wang *et al*. [58] implemented a Bidirectional Long Short Term Memory (Bidirectional-LSTM) approach for event extraction on Multi-Level Event Extraction (MLEE) corpus. Furthermore, for generalizing their approach they used BioNLP-ST-2009, 2011, 2013 corpora and achieved the F-scores more than 60% in the development set.

There is an increasing importance for biomolecular event applications and the current trends in biomedical relation extraction tasks, which uses ensemble learning methods and graph-based approaches [33, 42]. The motivations of our work integrate a Multiscale Laplacian Graph (MLG) kernel with a feature kernel as an ensemble model for the event extraction task. The challenge of the current study was the extraction of complex events using subgraph mining thereby gaining a deeper understanding of the biomolecular events. Kondor and Pan [53] first introduced MLG, and it was used to compare the structure in graphs simultaneously at multiple different scales. The objective of employing MLG in our event extraction is that it not only captures the topological relationships between the individual event nodes but also identifies the associations among the subgraphs for complex events.

The rest of the paper is organized as follows; Sect. 2 details the proposed materials and methods with a complete overview of the MLG model used in this study. Section 3 depicts the results and discussion followed by conclusions and future perspectives in Sect. 4.

2 Methods

The event extraction system presented in this study has three subtasks, namely (i) text preprocessing, (ii) event identification and (iii) argument detection. In text pre-processing, we applied general steps such as text preparation and cleaning, recognition of gene and protein mentions, dependency parsing of event sentences. In the event identification phase, we used two kernels, namely, a baseline feature-based kernel which uses token-based features, sentence-based features, parsing features, domain-specific features and the Multiscale Laplacian Graph (MLG) kernel, which uses the multilevel topological relationships between the event nodes as features. Both the feature-based kernel and the MLG kernel were combined using ensemble SVM for event identification. Finally, in the argument detection phase, we used lexico-syntactic patterns to detect arguments of the events. The overall schematic architecture of our event extraction pipeline has been depicted in Fig. 1 and each subtask is described detail in the following subsections.

In our methodology, we considered the nine most crucial events from BioNLP-ST [16–18], which are commonly used in existing studies. The nine types of events are merged into three main classes. The first five (Gene Expression, Transcription, Protein catabolism, Phosphorylation, Localization) had only one argument (theme: protein) and these events are called simple events. The second class of binding events involved more than one argument (two themes: proteins). Finally, the regulated events (Regulation, Positive regulation, Negative regulation) had two arguments: a theme and cause (event or protein).

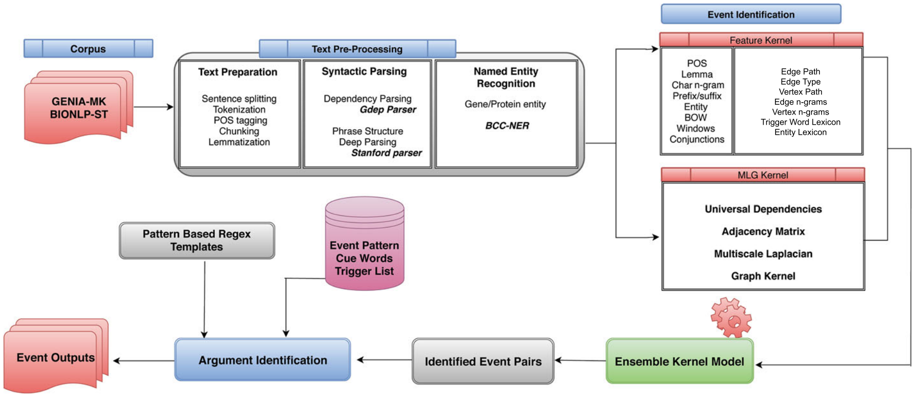


Fig. 1 Overall schematic architecture of the proposed event extraction system

2.1 Text pre-processing

2.1.1 Text preparation and cleaning

With a specific end goal to set up the corpus for extracting the events from it, the following preprocessing steps were carried out. They consisted of tokenization, sentence segmentation, POS tagging, lemmatization, and chunking. OpenNLP [59] was utilized for sentence splitting, tokenization, POS tagging, and chunking. Lemmatization was done by BioLemmatizer [60].

2.1.2 Dependency parsing

To provide information about grammatical relationships concerning two words extracted from a graph representation of the dependency relations in a sentence, we applied dependency parsing. The advantage of using dependency parsing is to find the grammatical relationships between two words and to find out the syntactic representation of a given sentence. A dependency relation is formalized as a direct grammatical relationship including two words (headword and dependent word) and a sentence is represented as a graph of dependency relations [61]. Dependency related features played an important role to extract the biomedical events. Here, we used two dependency parsers: the Stanford Dependency Parser (SDP) [62] is used to compute the universal dependencies and the GENIA Dependency Parser (GDep) [63], for the generation of the dependency graph of the sentence. Figure 2 depicts the dependency parse for a simple sentence. Here we can see that binary relations between common nouns such as *transcription*, *gene*, *activity* with adjectives and prepositions like *binding*, *in* and *c-jun* were identified. The given sentence explains Leukotriene B4 stimulates the transcription of genes c-fos and c-jun and activity AP-1 binding in human monocytes. The dependency parser identified *transcription, gene, Leukotriene, activity* as NN (noun, singular), *AP-1* as CD (cardinal number), and *monocytes* as NNS (noun, plural). The dependency parser also identified the grammatical relations within the sentence using amod (adjectival modifier), dobj (direct object), pobj (object of preposition), conj (conjuction), and prep (preposition).

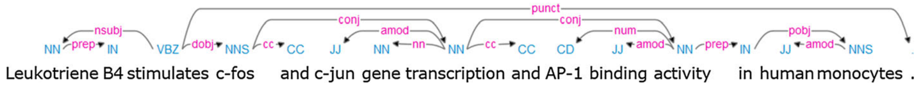

nsubj
prep
dobj
cc
conj
amod
nn
punct
conj
cc
num
amod
prep
pobj
amod
NN IN VBZ NNS CC JJ NN NN CC CD JJ NN IN JJ NNS .
Leukotriene B4 stimulates c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes .

Fig. 2 Dependency parsing for a simple sentence

2.1.3 Named entity recognition (NER)

The next step in our approach is the recognition of gene/protein mentions in the event sentences. To extract the events with high accuracy, named entities play an important role, since they came in the theme-cause role. NER is the process of detecting entities such as genes, proteins, diseases, species, RNA, cell, cell line from the text [64, 65]. BCC-NER [66], our in-house hybrid named entity tagger, was used to detect the gene and protein names automatically.

2.2 Event identification

Next, for event identification, we used an ensemble machine learning based classification approach with two kernels, namely feature-based kernel and MLG kernel. The feature-based kernel uses token-based features, sentence-based features, parsing features, and domain-specific features. The Multiscale Laplacian Graph Kernel (MLG) [53] uses the multilevel topological relationships between the event nodes as features. Both the feature-based kernel and the MLG kernel were combined using ensemble SVM [67] for event identification.

2.2.1 Feature-based kernel

In the baseline feature-based linear kernel, we used a total of 15 features broadly classified into four feature categories, namely token-based, sentence-based, parsing and domain-specific features which were employed successfully in a previous bio-event extraction task [68–70]. All 15 features are category wise grouped and illustrated in Table 1. The detailed feature representations for generating feature-based kernel model are clearly explained in Supplementary file S3.

2.2.2 Multiscale Laplacian Graph (MLG) kernel

Recently graph-based approaches for relation extraction are getting increased attention for their ability to capture both syntactic and semantic structures, thereby enabling deep understanding of the complex sentences such as bio-events and achieving state-of-the-art performances [41]. To improve the performance of the bio-event extraction task we employed the MLG kernel [53] along with the baseline feature-based kernel in our approach. The MLG kernel [53] is briefly introduced below and it is constructed based on two graph kernels, namely (i) Laplacian Graph kernel (LG), (ii) Feature space Laplacian Graph Kernel (FLG). The implementation of the MLG kernel is available at https://github.com/horacepan/MLGkernel.

Laplacian Graph (LG) Kernel: Consider graph G as the weighted undirected graph with vertex set $V = \{v_1, v_2 \dots v_n\}$ and the edge set E. The graph Laplacian [75] is a positive semi-definite matrix and it can be represented using adjacency matrix A and weighted degree matrix D. The Laplacian matrix of the graph can be expressed using the notation $L = D - A$.

Table 1 Category wise features used in feature-based kernel

Feature category	Description	Features	Example
Token-based features	Features that are based on tokens and expect to catch specific knowledge, for every token specifically linguistic, orthographic and morphological characteristics [73]	POS (Part-of-speech)	**POS**: Detect the grammatical role of the word in a given sentence **Example:** NP-FOXP3 NNP-Represses NNP-retroviral NNP-Transcription PP-by CC-Both NNP-NF-kappaB CC-and NNP-CREB NNPS-Pathways
		Lemma	**Lemma**: *EBV Latent Membrane Protein 1 'Activates'Akt, NFkappaB, and Stat3 in B Cell Lymphomas* **Example:** *Here the word* ***activates*** *was converted into* ***activate*** *by the implementation of lemmatization*
		Char n-gram	**Char n-gram:** subsequence of n characters from the given token **Example:**Promotors Pro / mot/ ors….rom, omo, oto, tor etc., n-gram size = 3
		Prefix/suffix	**Prefix/suffix**: The two-character prefix ***up*** was used to denote the trigger word ***upregulation***
		Orthographic features	Orthographic features were used to describe the presence of capitalization, punctuation, and numeric or special characters **Example:** *Reactive oxygen intermediate-dependent NF-kappaB activation by interleukin-1beta requires 5-lipoxygenase or NADPH oxidase activity* In the above example sentence, the orthographic and linguistic features (**Ex: Mixcaps-*NF-kappaB,* numeric - interleukin-1beta, *All caps - NADPH***) are used to describe protein names [71]

Table 1 continued

Feature category	Description	Features	Example
Sentence based features	Sentence level features were used to identify the common characteristics of the sentence [74]. It includes a number of tokens (windows and conjunctions of features), frequency of recognized named entities (Entity counts), and number of words (bag-of-words) for a particular sentence	Entity counts	**Entity count**: Count the number of entities (proteins) in the sentence **Example:** ***Foxp3*** *Represses Retroviral Transcription by Targeting Both* ***NF-kappaB*** *and* ***CREB*** *Pathways* Number of entities in the sentence: 3 (Foxp3, NF-kappaB, CREB)
		BOW counts	**BOW counts**: This feature is used to count how many times each word appears in the document **Example:** *Foxp3 Represses Retroviral Transcription by Targeting Both NF-kappaB and CREB Pathways* BOW count of each word in the above sentence against BioNLP 2011 corpus is as follows Foxp3-202, Represses- 9, Retroviral-19, Transcription-1815, by-2472, Targeting-14, Both-521, NF-kappaB-424,and-6539,CREB-120,Pathways-155 counts etc
		Windows or conjunctions of features	**Windows or conjunctions of features:** Higher-level relations among tokens and separated features can be built up through windows or conjunctions of features, reflecting the local context of every token, it comprised of POS, lemmas, and n-grams extracted from the word around the target (trigger word, protein) token

Table 1 continued

Feature category	Description	Features	Example
Parsing features	The token and sentence-based features gave the local analysis of the sentence [69] but parsing features were used to get the global information about relations between the tokens of a sentence. Finding grammatical relationships, syntactic representation between the two words (trigger word, protein), parsing features play an important role [70]. Moreover, parsing features gave the hierarchical relationships between the trigger words and proteins	Edge path	**Edge path:** Path of edge labels between two tokens **Example:***Foxp3 Represses Retroviral Transcription by Targeting Both NF-kappaB and CREB Pathways* (we calculate the path between the two tokens: Foxp3 and retroviral NMOD_Foxp3, PMOD_Represses, NMOD_Retroviral) The Edge Path feature, Foxp3 and retroviral NMOD_Foxp3, PMOD_Represses, NMOD_Retroviral, the aim was to finding grammatical relationships, syntactic representation between the two words (trigger word, protein) and to detect hierarchical relationships between the trigger words and proteins. As followed in the dependency parsing approach, a matrix is created with words (trigger word, protein) representing the column and if a path is existed between any two of these, all the keywords participating in that path are labeled as 1 otherwise 0. This approach is adopted for all the paths extracted from the graphs created
		Edge type	**Edge type:** Represents the type of the edge path based on its size and first edge label The path between Foxp3 to transcription: length = 3

Table 1 continued

Feature category	Description	Features	Example
		Vertex path	**Vertex path: path exists between two token features in this case vertexes** **Example:**Foxp3 Represses Retroviral Transcription by Targeting Both NF-kappaB and CREB Pathways “by-Targeting-Both-NF-kappaB-and-CREB”, by using enhanced Universal Dependencies, nmod:by (Transcription-4, NF-kappaB-8) nmod:by (Transcription-4, Pathways-11) Here the nmod (nominal modifier) is a noun functioning. The nmod relation used to represent prepositional complements, i.e., It holds between the noun/predicate modified by the prepositional complement and the noun introduced by the preposition. An example has been given as “Foxp3 Represses Retroviral Transcription by Targeting Both NF-kappaB and CREB Pathways” in page no:10 Table 1. In the above example sentence the vertex NF-kappaB, pathways are targeted by the enhanced universal dependencies and we have taken transcription is an event node that targeted the two entity nodes NF-kappaB and CREB pathways
		Edge n-grams	**Edge n-grams:** n-grams of edge labels between two tokens. (NMOD_PMOD,PMOD_NMOD:Foxp3_Represses, Represses_retroviral)
		Vertex n-gram	**Vertex n-grams:** n-gram of features of tokens (vertexes) between two tokens (Transcription by, by_target (consider 2 grams and lemmas as features.))

Table 1 continued

Feature category	Description	Features	Example
Domain-specific features	To further optimize the event extraction approach, domain-specific features were used to extract the features from external resources such as lexicons. Here we used trigger word lexicon and gene/protein name lexicons to detect the presence of the trigger words or entity. We employed trigger word lexicon from TrigNER [70] and gene/protein lexicons from UniProtKB [72] to further improve our event extraction methodology	Trigger word lexicon [70] Entity lexicon [72]	**Trigger words:** Active, activated, inhibited, etc **Entities:** IL-1, IL-4, CpG, etc

The LG kernel of two graphs (G_1, G_2) can be defined by the following equation.

$$k_{\mathrm{LG}}(G_1, G_2) = \frac{\left|\left(\frac{1}{2}S_1^{-1} + \frac{1}{2}S_2^{-1}\right)^{-1}\right|^{1/2}}{|S_1|^{1/4}|S_2|^{1/4}} \tag{1}$$

where $S_1 = L_1^{-1} + \lambda$ I, $S_2 = L_2^{-1} + \lambda$ I.

The L_1^{-1}, L_2^{-1} are the inverse of the graph Laplacian and I is the identity matrix with parameter λ, these are used to obtain the similarity between the graphs G_1, G_2.

Feature Space Laplacian Graph kernel (FLG): FLG kernel was used to compare the structure of the subgraphs in a single scale. FLG unites the information attached to the vertices with the graph Laplacian. The advantage of employing the FLG kernel is to transform the vertex space variables $a_1, a_2 \ldots a_n$ into feature space variables $b_1, b_2 \ldots .b_n$, where $b_i = \sum_j t_{i,j}(a_j)$ and each $t_{i,j}$ only depend on j during local and reordering the invariant possessions of vertex $\mathrm{v_j}$ and the resulting kernel should be permutation invariant. Vertex space variables are the input variables that can be used to transform graph vertex as the feature vertex. Consider G_1, G_2 as the two graphs with regularized Laplacians L_1 and L_2, and we define the parameter $\lambda \geq 0$ and ($\Phi_1,\ldots,\Phi_m$) is a collection of m local vertex features and they define the feature mapping matrices in the FLG. The FLG kernel is defined as follows.

$$k_{\mathrm{FLG}}(G_1, G_2) = \frac{\left|\left(\frac{1}{2}S_1^{-1} + \frac{1}{2}S_2^{-1}\right)^{-1}\right|^{1/2}}{|S_1|^{1/4}|S_2|^{1/4}} \tag{2}$$

where $S_1 = \mathrm{U}_1 L_1^{-1} U_1^T + \lambda$ I, $S_2 = U_2 L_2^{-1} U_2^T + \lambda$ I

Here U_1 and U_2 are the feature mapping matrix, L_1 and L_2 are the Laplacian matrix and I is the identity matrix with parameter λ and transpose U_1^T ,U_2^T . The major limitation of the FLG kernel is that it cannot consider graph structure at multiple different scales which paved the way for the MLG kernel. The FLG kernel acts as the key component in the MLG kernel and it is applied recursively for the construction of MLG.

Multiscale Laplacian Graph (MLG) Kernel: The MLG kernel for a graph (G) can be computed as follows:

(i) The graph (G) is divided into a large number of smaller subgraphs, and the FLG kernel is computed between any two subgraphs for the similarity calculation in single scale.
(ii) A new kernel (FLG) is calculated between the vertices by placing the extracted subgraphs to a random vertex of the graph G.
(iii) Finally, a new FLG kernel is computed between the large subgraphs of the graph (G) based on step ii and this process is repeated L (multiple scales) times.

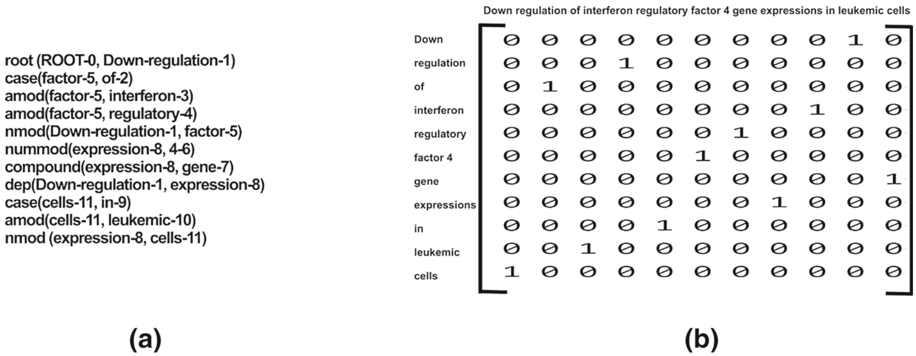
The MLG kernel thus constructed as follows:

Consider G as the graph with vertex set V, and compute the kernel k as a positive semi-definite kernel on the vertex set V. For each vertex (v) in the vertex set V ($v \in V$) we have a nested sequence of L neighborhoods.

$$v \in N_1(v) \subseteq N_2(v) \subseteq \cdots \subseteq N_L(v) \subseteq V \tag{3}$$

Consider $G_l(v)$ as the corresponding subgraph for each $N_l(v)$. From the above equation, the Multiscale Laplacian subgraph (MLS) kernel can be defined by calculating multiple FLG kernels for vertex set V as ($k_1 \ldots k_L$: V × V → R).

$$k_1(v, v') = k_{\mathrm{FLG}}^{k}(G_1(v), G_2(v')) \tag{4}$$



root (ROOT-0, Down-regulation-1)
case(factor-5, of-2)
amod(factor-5, interferon-3)
amod(factor-5, regulatory-4)
nmod(Down-regulation-1, factor-5)
nummod(expression-8, 4-6)
compound(expression-8, gene-7)
dep(Down-regulation-1, expression-8)
case(cells-11, in-9)
amod(cells-11, leukemic-10)
nmod (expression-8, cells-11)

Down regulation of interferon regulatory factor 4 gene expressions in leukemic cells

$$\begin{array}{l}
\text{Down} \\ \text{regulation} \\ \text{of} \\ \text{interferon} \\ \text{regulatory} \\ \text{factor 4} \\ \text{gene} \\ \text{expressions} \\ \text{in} \\ \text{leukemic} \\ \text{cells}
\end{array}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

Fig. 3 Universal dependencies and adjacency matrix for the example sentence (a) universal dependencies, (b) adjacency matrix

k_1 is the FLG kernel (k^{k}_{FLG}) generated from the base kernel k. Here, the base kernel k is used to boost the FLG to multi-scale kernel.

$$k_l\left(v, v'\right) = k^{k_{l-1}}_{\mathrm{FLG}}\left(G_l(v), G_l\left(v'\right)\right) \tag{5}$$

where $l = 2, 3 \ldots L$, and k_l is generated from k_{l-1} kernel.

Let G be a set of graphs as a chance to be an accumulation of graphs with the end goal that all their vertices are members of an abstract vertex space V supplied with a symmetric positive semi-definite kernel $k : V \times V \rightarrow R$. Assume that the MLS kernels k_1,…,k_L are characterized in Eqs. 4 and 5 both for pairs of subgraphs inside the same graph and crosswise over pairs of different graphs. Now the MLG kernel can be structured as follows

$$k(G_1, G_2) = k^{\mathrm{LG}}_{\mathrm{FLG}}(G_1, G_2) \tag{6}$$

In this study to implement the MLG kernel, we generated Universal Dependencies (UD) along with the adjacency matrix of the bio-event sentences.

Universal dependencies: We applied Stanford parser for generating UD of the sentences [62]. The grammatical relations of UD are described in a hierarchy, rooted in the most generic relation dependent. In this study, we applied UD in all event sentences to extract the typed relation across the sentence, especially with trigger words and entities.

Adjacency matrix: The generated UD of biomedical event sentences was used to create an adjacency matrix, to represent the association between words. An example UD generated and corresponding adjacency matrix for a sample sentence (*PMCID: 1310901*) is shown in Fig. 3a, b, respectively.

Subgraph mining

In the MLG kernel, the subgraph mining process was essential to scale the event sentences at multiple levels. The aim of this graph kernel is to find the local structures that are critical at specific position of the graph and find global property that roughly summarizes the graph. In order to do so, MLG kernel is defined as a graph kernel that can consider structure at multiple scales, by comparing graphs by subgraphs recursively. The underlying procedure is that, two graphs are compared by subgraphs, in the next iteration two subgraphs are compared by smaller subgraphs and so on. The MLG kernel uses node features to capture the global structure and induced feature vectors by similarity scores for comparing structures at multiple scales. Recursive approach compares the same subgraph pairs multiple times by calculating the similarity scores on smaller neighborhood. In this study, we created the graph

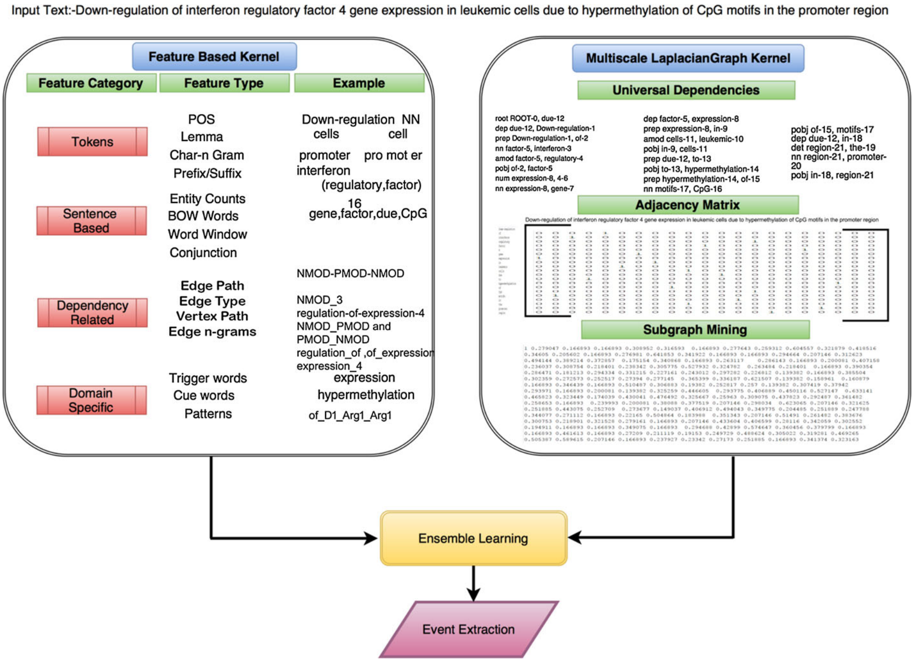


Fig. 4 An example of an ensemble classification pipeline of the two kernels

using Universal Dependencies (UD) along with adjacency matrix. The subgraph mining was carried out using the following procedure. (i) First, assign the node degree to the entire graph-structured event sentence. (ii) Construct the subgraph from the large graph. (iii) Design a larger subgraph for the event sentence. (iv) Assign the low-rank approximation approach to entire subgraphs and each larger subgraphs.

2.2.3 Ensemble classification

In the biomedical domain, ensemble classification plays a vital role in improving overall performance for tasks such as article classification [76, 77] and relation extraction [6, 78]. SVM with an ensemble learning approach productively learns multiple training models through lowest time complexity. In the EnsembleSVM [67], bootstrapping strategy was employed to repeatedly learn the training models and aggregates the multiple training instances into the single predicted model. In this study, we employed EnsembleSVM [67] to generate the ensemble models for feature-based linear kernel and MLG kernel and merge them to a single classification model to efficiently categorize the events. Using EnsembleSVM we created models on bootstrap subsamples and trained ensembles of SVM models for feature based and MLG kernel, respectively. Figure 4 depicts a detailed explanation of the ensemble classification pipeline of our approach.

Ensemble classification of our approach described in Eq. 7:

$$E_k = F_k + G_k \tag{7}$$

Here,E_k,F_k and G_k were the kernel models in our classification problem. Using the “validation set”, we tuned various parameters using the grid search method in our model generation. In the features section, char n-gram was set to 3 and prefix/suffix feature assigned as two-character. In the MLG kernel model, parameters were optimally generated and finally set as the *radius* to 3, *levels* to 4, *eta* to 0.1, *gamma* to 0.01 and *threads* to 32. The tree value parameter *grow* was set to 1 to grow by leaf radius. This is for allowing the subgraphs to double in size at each level. We kept all these parameters to their default values during the model development.

2.3 Argument detection

After the identification of events and triggers, the next step is to extract arguments, which describe the events. To extract arguments for the events from the text efficiently and accurately, we used the lexico-syntactic pattern-based approach with semantic role labeling [79] which is briefly introduced below.

2.3.1 Lexico-syntactic pattern and semantic role-based rules engine

Lexico-syntactic patterns [80] are generalized linguistic structures for extracting related concepts and relationships between concepts from the text. Here the trigger words and propositions (synonym, subject, and verb) were the concepts and relationships to detect the event arguments. Lexico-syntactic patterns were used to structure the ontology of the words. Motivated by the work of Hung et al. [79], we employed lexico-syntactic patterns to identify {THEME, CAUSE} of the events. In the current study to identify arguments from the events, a combination of lexico- syntactic patterns and semantic matching were performed through three steps, namely contextual patterns, semantic role labeling, and event-specific argument structure, respectively. The list of bio event cues and trigger word list were used to match the arguments using pattern matching and role labeling. In the event-specific argument structure phase, post-processing rules were incorporated such as emphasizing event certainty and co-reference mentions. A detailed description of our lexico-syntactic pattern-based rule engine is depicted in Fig. 5. A brief explanation about each step incorporated in the rule engine with an example is discussed below.

Contextual patterns

Contextual patterns (CP) utilize domain specific information such as a trigger word list and tagged entities to annotate possible event arguments. The contextual patterns were employed with the following two components: *subclass* and *complex*. *Subclass* was utilized to detect and annotate the trigger word list and tagged entities using the pattern keyword (*VP list, VP, NP)* from the dependency parsed sentences. These patterns were also used to detect the *prepositions* (to, belong, with, without, etc.) between the trigger words and proteins. Tagged entities are represented in the sentence as 'protein 1' and 'protein 2' etc. For example: *interact with Protein 1 and Protein 2. Complex Patterns* were employed to identify the verb keywords which indicate multiple arguments of the same events. For example, protein1 *interacts with* protein 2 which *catalyzes* protein 3 and *causes* protein 4 *downregulation*. The above sentence contains multiple events which is represented by the cue words '*catalyzes'*, '*interacts with'*, '*cause*'. A full example of contextual pattern identification is shown below. For example: (PMID:9973520)

Rule 1:

Scenario: *Identifying the subject and verb in a given sentence.*

Original sentence: *Cross-linking of CD44 on rheumatoid synovial cells up-regulates VCAM-1*

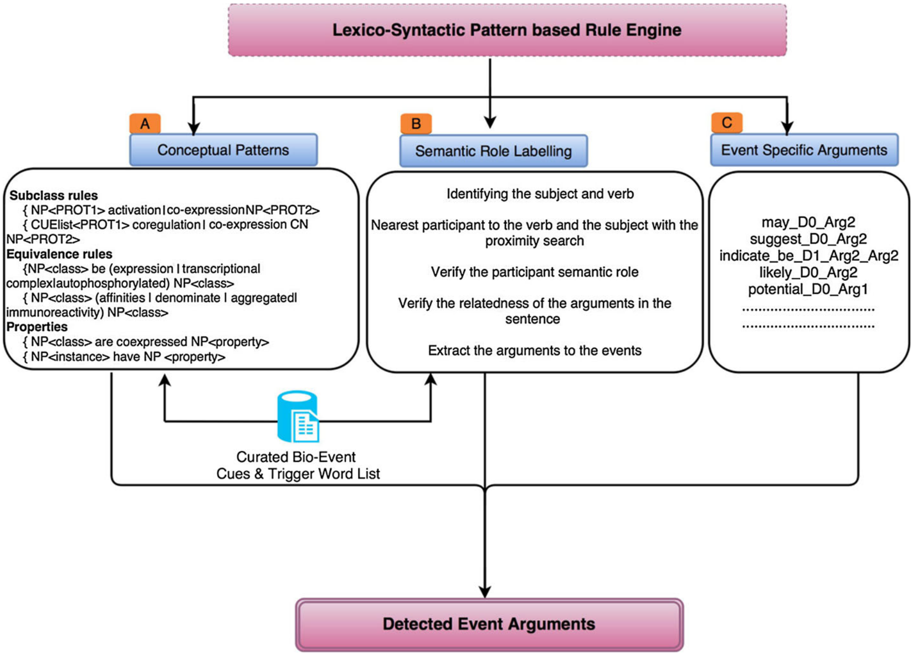

Lexico-Syntactic Pattern based Rule Engine
A
Conceptual Patterns
B
Semantic Role Labelling
C
Event Specific Arguments
Subclass rules
{ NP<PROT1> activation|co-expressionNP<PROT2>
{ CUElist<PROT1> coregulation| co-expression CN NP<PROT2>
Equivalence rules
{NP<class> be (expression | transcriptional complex|autophosphorylated) NP<class>
{ NP<class> (affinities | denominate | aggregated| immunoreactivity) NP<class>
Properties
{ NP<class> are coexpressed NP<property>
{ NP<instance> have NP <property>
Identifying the subject and verb
Nearest participant to the verb and the subject with the proximity search
Verify the participant semantic role
Verify the relatedness of the arguments in the sentence
Extract the arguments to the events
may_D0_Arg2
suggest_D0_Arg2
indicate_be_D1_Arg2_Arg2
likely_D0_Arg2
potential_D0_Arg1
Curated Bio-Event Cues & Trigger Word List
Detected Event Arguments

Fig. 5 Lexico-syntactic and semantic role based rule engine for argument detection phase

After applying CP: *NP → Cross-linking VP → up-regulates etc.,*
Contextual patterns identify and annotate the possible trigger words and entities by utilizing the trigger word list in the sentence, which will be processed further by applying semantic role labeling techniques described below.

Semantic role labeling
Semantic role labeling (SRL) is a process in natural language processing to determine the relationship between the verb and syntactic structure of a sentence [79]. In our approach, semantic role labeling was used to search and determine the association between protein entities and trigger words in a sentence. It involves the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles. Here we have taken the sentence (A), Verb (VP), modal verb or preposition (M), participants in the sentence (P) and S as the subject of the sentence. From the above steps, we derived a semantic role labeling approach for our event arguments construction process.

For example, in the following sentence (PMID:9973520) "*Cross-linking of CD44 on rheumatoid synovial cells up-regulates VCAM-1*" the trigger word was "*up-regulates*", the arguments were *CD44, VCAM-1 and rheumatoid synovial cells* and they participated in the event *positive regulation.* This was identified by applying a set of rules; the procedure of the same has been given in Table 2 in detail.

Event-specific argument structure
After extracting event arguments using SRL based patterns, we incorporated two post-processing rules as event-specific argument structures to raise the performance of our event argument detection approach. Event-specific argument structure was used to differentiate simple and complex events that specify the arguments directly or indirectly in the tagged sentences.

Table 2 Rules used for detecting the arguments of the events

Rule	Original sentence (A)	After Conversion ($$ denotes the starting and ending of sentence)
Rule 2: Nearest participant (P) to the verb (VP) and Subject (S) with the proximity search (The participant nearest to the subject(S) of that verb(VP), are considered to be arguments of the event for which that particular verb(VP) is a trigger word	Cross-linking of CD44 on rheumatoid synovial cells up-regulates VCAM-1	**$$** (S- Cross-linking, M - of, P - CD44, M-on, NP- rheumatoid synovial cells, VP-up-regulates, P – VCAM)**$$** Here the **P- CD44** was nearest to the S-Cross-linking, and **P-VCAM-1** was nearest to the **VP - up-regulates**
Rule 3: Determine the participant (P) semantic role (The semantic role is the underlying relationship that a participant has with main verb in a clause)	Cross-linking of CD44 on rheumatoid synovial cells up-regulates VCAM-1	**$$** (S- Cross-linking, M - of, P - CD44, M-on, NP- rheumatoid synovial cells, VP-up-regulates, P – VCAM)**$$** Here the **P- CD44** relates in the event "**positive regulation"** by the use of **VP-up-regulates**
Rule 4: Find out the relatedness of the participants (P) in the sentence (The arguments and phrases such as NPs and VPs are used to retrieve the true positive arguments)	Cross-linking of CD44 on rheumatoid synovial cells up-regulates VCAM-1	**$$** (S- Cross-linking, M - of, P - CD44, M-on, NP- rheumatoid synovial cells, VP-up-regulates, P – VCAM) **$$** Here the participants **CD44, VCAM-**1 participated in the event "**positive regulation"** by the use of **NP- rheumatoid synovial cells**, **VP-up-regulates**

(i) Searching for a connective pronoun such as "it" in the sentence which indicates the entity names (Protein) in the below example.

Examples of the generated rules: *(Here D0, D1-Dependencies, ARG1, ARG2-Arguments of the particular word in the sentence)*

it_D0_Arg1_Arg2
both_D0_ D1_Arg1_Arg2
that_ D0_D1_Arg2_Arg1

Example: PMID: 10209041

Expression of GrpL is restricted to hematopoietic tissues and ***<Keyword>*** *it* ***</Keyword>*** *is distinguished from Grb2 by having a proline-rich region.*

In the above example the pronoun *'it'* denotes the protein *GrpL,* and it participated in the event *'Expression'*.

(ii) Searching for specific keywords such as 'certainly', 'highly', 'confirm', which were co-mentioned with trigger words 'activation' or 'up-regulation' so that event-specific meaningful sentence can be identified rather than a generalized one. This is also used to identify specific trigger words, which describe the event accurately from multiple trigger words in the same sentence.

Table 3 Corpus statistics (Abs—Abstract, Full—Full text articles)

Corpus	Statistics		
BioNLP-ST	Training	Development	Test
BioNLP-2009 [16]	Abs-800	Abs-150	Abs-260
BioNLP-2011 [17]	Abs-800, Full-5	Abs-150, Full-5	Abs-260, Full-4
BioNLP-2013 [18]	Full-10	Full-10	Full-14
GENIA-MK [21]	1000 Abs		

Examples of the generated rules:

highly _D0_Arg1, probably_D0_D1_Arg1, certainly _D0_D1_Arg2, confirm _D0_Arg1 etc.,

<Keyword>confirm</Keyword>*<D0>that</D0><ARG1>binding</ARG1><D1>of </D1>endogenous <ARG2>NFkappaB</ARG2> and <ARG3>AP1<ARG3>.*

Example: PMID: PM9190901

We ***<Keyword>confirm</Keyword>*** *that binding of endogenous NFkappaB and AP1 is induced following PMA/ionomycin treatment of T cells.*

In the above example, the keyword '*confirm'* described the certainty of the event 'binding'.

The analysis of training data was used to makeup the lexico-syntactic pattern-based rule engine to detect the participating themes in the events. We developed a pattern matching module using Java Regex [22–24] coupled with the above process to detect the arguments in the event classes.

3 Results and discussion

3.1 Dataset

For the first time, BioNLP-ST-2009 [16] introduced three tasks based on the GENIA corpus [20] for the detection of core events, recognition of event arguments and negation/speculation detection. In BioNLP-ST-2011 [17], the tasks were expanded with resources to capture more text and event types. In BioNLP-ST-2011, the GENIA Event extraction (GE) task has been kept and augmented with three focused event tasks, namely (i) epigenetic and post-translational modification (EPI), (ii) bacteria biotope (BB) and bacteria interaction (BI) and (iii) infectious diseases (ID) [17]. Application domains were further expanded in BioNLP-ST-2013 [18] while keeping the GE and BB; the additional tasks were cancer genetics (CG), gene regulation ontology (GRO), and pathway curation (PC).

To assess the performance of our approach, we employed four different corpora which includes three corpora from BioNLP-shared task (BioNLP-09 [16], BioNLP-11 [17], BioNLP-13 [18]), and one another standard corpus, namely GENIA-MK (Meta-knowledge) [21] which is currently available and widely used for event extraction tasks. All the four corpora were used to train and test the models of our approach. The corpus statistics of all three BioNLP-ST corpora and the GENIA-MK corpus are represented in Table 3.

3.2 Evaluation metrics

Evaluation of our event extraction system was performed based on standard evaluation metrics precision (P), recall (R), and F-Score (F). The shared task online evaluation server was used to perform the evaluation of the BioNLP-ST (2009, 2011, 2013). The results reported in our system are based on Approximate span matching and Approximate string matching evaluation measures. For the GENIA-MK corpus evaluation, we used 10-fold cross validation. In the 10-fold cross validation the GENIA-MK corpus was divided into 10 subsets. Every run, 90% of the data was used as the training set, and the remaining 10% was used as the test set.

3.3 Evaluation results

We trained and tested our approach on BioNLP-ST 2009, 2011, 2013 and GENIA-MK corpus with the Feature-based linear kernel, MLG kernel, and Ensemble kernel. Following training and testing, approaches were carried out to assess the performance of our approach.

In Table 4, first, we implemented the ensemble feature-based approach on the BioNLP-ST-2009 corpus. By analyzing Table 4 feature-based approach results in high precision and low recall. Next, we deployed the ensemble MLG kernel-based approach to the corpus, and it results in high precision and high recall and moderately increases the F-score. Finally, we combined both ensemble kernels, which takes the benefits of both feature-based and MLG kernel-based output models and attained the comparative F-score. Likewise, we applied the above methods in BioNLP-ST-2011 and BioNLP-ST-2013 corpus.

In Table 5, we implemented the same approach on the GENIA-MK corpus. Experimental results show that our approach attained the best results compare to the BioNLP-ST corpora. Figure 6 depicts the Receiver Operating Characteristic (ROC) curve of the three kernels for all four corpora.

To classify the events individually, every event type needs a variety of features to reflect the diverse context and linguistic characteristics. For example, compared to the events such as *gene expression, transcription, localization, the regulation* events need more token-based, concept based and syntactic information. By the implementation of a feature-based approach in our study, we properly modeled the higher complexity associated with their phrasal and linguistic contexts and consequently prepared our model to identify the individual events. Next, the feature-based approach was coupled with MLG kernel that takes advantage of both feature-based and graph-kernel based approaches and generated state-of-the-art performances in the extraction of individual classes of events. Table 6 shows results for individual classes of events in the four corpora by employing our ensemble approach.

Next, we compare our approach with other state-of-the-art approaches developed on the BioNLP-ST 2009, 2011, 2013 and GENIA-MK corpora. Comparisons show that our proposed approach performs better than other state-of-the-art approaches. Tables 7 and 8 show the comparisons.

3.4 Discussion

In our methodology, we implemented a Feature-based linear kernel, MLG kernel, and lexico-syntactic pattern-based approaches to extract biomedical events with unique steps. Some interesting findings encountered from our approach are discussed below. The baseline feature-based linear kernel captured grammatical, syntactical, morphological, orthographical, and sentence level global information successfully. Morphological and orthographical features

Table 4 Results on BioNLP-ST Corpora

Approach	Dataset (BioNLP-ST)								
	BioNLP-ST-2009			BioNLP-ST-2011			BioNLP-ST-2013		
	P (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)
Ensemble feature-based linear kernel	64.75	58.50	61.46	60.82	56.15	58.40	57.12	53.25	55.11
Ensemble MLG kernel	65.92	59.75	62.68	62.12	58.75	59.80	58.35	54.75	56.49
Ensemble feature-based linear kernel + Ensemble MLG	66.75	61.12	63.79	63.45	60.12	61.74	60.55	56.23	58.30

Table 5 Results on GENIA-MK Corpus

Approach	*P* (%)	*R* (%)	*F* (%)
Ensemble feature-based linear kernel	58.12	54.25	56.12
Ensemble MLG kernel	60.13	58.35	59.22
Ensemble feature-based linear kernel + Ensemble MLG kernel	64.12	59.25	61.58

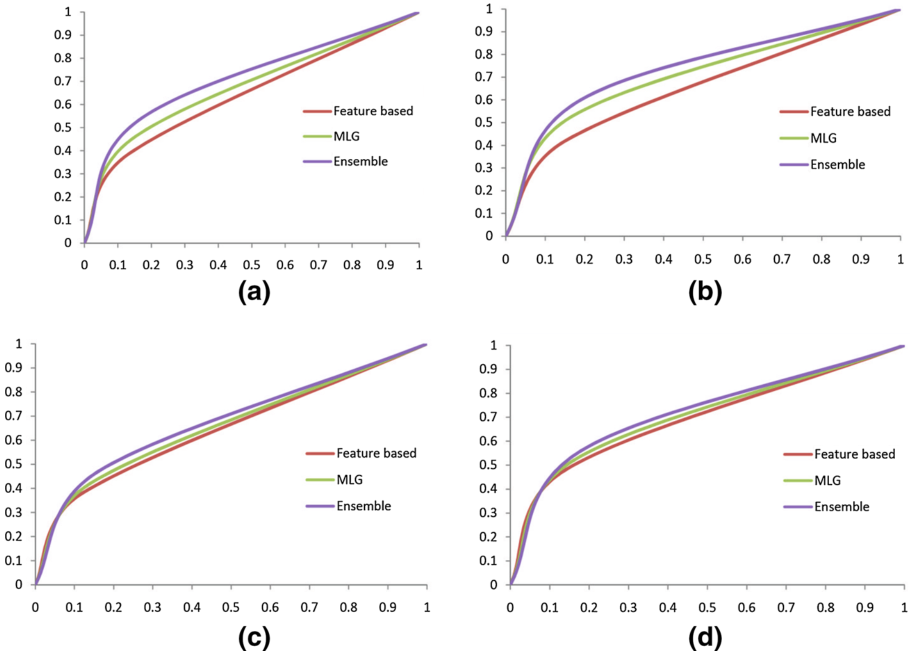


Fig. 6 ROC plotting results for four corpora **a** BioNLP-ST- 2009, **b** BioNLP-ST-2011, **c** BioNLP-ST-2013, **d** GENIA-MK

were used to describe the structure of the word in a given sentence. Parsing features were employed to discover the grammatical and syntactical expressions of the event sentences. Packing these above features and methodologies in the feature-based linear kernel gave the perfect baseline to extract the events from the biomedical literature.

The MLG kernel was used to compare the structure of the graph at multiple different scales. Mining subgraphs is an important phase in the MLG kernel because each generated subgraph will be compared by its constituent sub-subgraphs. MLG kernel first accepts a universal dependency structure, in which a direct dependency relationship path between the trigger words and named entities. MLG kernel combines baseline graph Laplacian kernel with feature representations originating from nested neighborhoods. Finally, MLG kernel considers both overall and local graph structures to learn similarities at multiple different levels. By considering all this we believe that by employing MLG kernel, our system was not only able to capture the topological relationships between the individual event nodes but also identifies the topological relationships between the subgraphs.

Example: PMCID 1310901

Table 6 Results for individual classes of events on BioNLP-ST 2009, 2011, 2013 and GENIA-MK

Event classes	Dataset											
	BioNLP-ST-2009			BioNLP-ST-2011			BioNLP-ST-2013			GENIA-MK		
	P (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)	*P* (%)	*R* (%)	*F* (%)
Gene expression	68.12	62.75	65.32	66.72	60.21	63.29	64.12	58.25	61.04	68.15	64.25	66.14
Transcription	62.12	56.32	59.07	64.12	58.75	61.31	62.22	56.25	60.88	60.12	56.75	58.38
Protein catabolism	63.19	57.12	60.00	64.30	58.17	61.08	60.27	56.12	58.12	64.12	60.25	62.12
Localization	64.12	58.72	61.30	60.12	54.72	57.29	59.75	53.12	56.24	63.15	59.25	61.13
Phosphorylation	62.12	56.75	59.31	63.12	57.28	60.05	62.19	56.72	59.32	60.12	56.75	58.38
Binding	58.12	52.00	54.88	55.13	49.12	52.00	53.12	46.75	49.73	59.75	53.12	56.24
Regulation	49.75	43.12	46.20	45.12	39.72	42.54	43.16	39.72	41.36	50.12	46.75	48.37
Positive regulation	50.72	46.21	48.36	46.75	40.12	43.18	48.25	41.75	44.80	46.12	39.75	42.69
Negative regulation	53.16	45.72	49.16	47.12	40.27	43.42	49.12	42.72	45.70	50.12	46.85	48.43

Table 7 Comparative analysis on the BioNLP-ST CORPORA (BioNLP-ST-09, BioNLP-ST-11, and BioNLP-ST-13) based on F-score (%)

System	Approach	BioNLP-ST-2009	BioNLP-ST-2011	BioNLP-ST-2013
Bijorneet al. [33]	SVM-multiclass	51.95	–	–
Miwa et al. [81]	SVM-Multiclass + Shortest path features	53.29	–	–
Riedel and McCallum et al. [55]	Stanford event parser + dependency trees + joint prediction	–	56.05	–
Riedel et al. [36]	Stanford event parser + dependency trees	–	57.46	–
Hakala et al. [35]	Re-ranking-SVM	–	–	50.97
Bijorne and Salakoski et al. [34]	SVM + task specific event centric rules	–	–	50.74
Bui et al. [82]	Dictionary + pattern matching rules	–	–	50.68
Liu et al. [39, 40]	Exact Subgraph Matching + Approximate Subgraph Matching (ESM/ASM) approaches	–	51.12	–
Liu et al. [41]	(ESM/ASM)+DSM(Distributional Similarity Model)	–	–	55.09
Munkhdalai et al. [56]	Semi-supervised learning+ self-training in significance space (STSS)	–	54.30	–
Wang et al. [30]	multiple distributed representation method+ context-based word embedding	59.94	55.20	50.12
Li et al. [57]	dual decomposition method+ dependency parse based rich features	56.09	–	53.19
Bijorne and Salakoski [83]	Convolution neural networks	57.84	58.10	53.00
Our approach	Ensemble Feature-based Linear kernel + Ensemble MLG Kernel	63.79	61.74	58.30

Table 8 Comparative analysis on the GENIA-MK Corpus in terms of F-score

System	Approach	*F* (%)
Miwa et al. [84]	SVM + Meta-knowledge information	58.20
Our approach	Ensemble feature-based Linear kernel + Ensemble MLG Kernel	61.58

Sentence: *Downregulation of interferon regulatory factor 4 gene expression in leukemic cells*

In the above example, words in the sentences were converted to universal dependencies and then to the adjacency matrix. The MLG kernel first assigns the node degrees based on UD and adjacency matrix to the graph generated for the sentence. In our case, in this example, words like *expression* and *factor* were assigned with high node degree.

In general, the graph structure is captured at multiple scales in MLG. This is achieved by increasing the depth of the neighborhood vertices in the graph. In addition, MLG focuses on capturing the neighborhood similarity among the vertices and uses this similarity score to induce the feature vectors. The current study exploits the above technique in which the biomolecular event sentence is searched at multiple scales for finding the relations between events and the target proteins using the graph generated from the corresponding adjacency matrix of the sentence. An interesting connection to be noted is that the cue words like *gene and expression*, *regulation and factor*, *leukemia and cells* were connected in the graph. In the following steps, a subgraph mining from the sentence graph followed by the building of larger subgraphs was performed. As a result of this step, words like *interferon, regulatory, factor, gene, expression* are added into a single subgraph. So, we strongly believe that our subgraph mining based MLG kernel played an important role in capturing the key information about the biomedical event sentences.

The association among the subgraphs for complex event extraction using MLG kernel is represented in Fig. 7 for a sample sentence from PMID 1335418. From the sentence, the MLG kernel first detects the small subgraphs in level one as entity names and event trigger words (For example, *cAMP* and *accumulation*). In level two, the kernel identifies the relationship between the trigger word and the corresponding proteins by accumulating multiple subgraphs (*activation, cells, jkat, protein, kinase, cells*). Finally, in level three the larger subgraphs were mined, thereby identifying the complex event. The repeated subgraph mining process was done until the low-rank approximation was observed to improve the classification accuracy.

Example: PMID: 1335418

We have earlier found that in Jurkat cells activation of protein kinase C (PKC) enhances the cyclic adenosine monophosphate (cAMP) accumulation induced by adenosine receptor stimulation or activation of Gs.

Next, in argument detection, we employed lexico-syntactic based semantic role labeling and contextual pattern-based rules to extract the event arguments efficiently. Lexico-syntactic patterns were used to detect domain-specific ontology-based concepts and relationships effectively. In the event extraction task, lexico-syntactic patterns with semantic role labeling process require significantly less time to compare normal lexico-syntactic patterns. The examples were illustrated in detail in the methods Sect. 2.3.1. A few interesting advantages of using lexico-syntactic patterns to event argument detection are illustrated in the following examples.

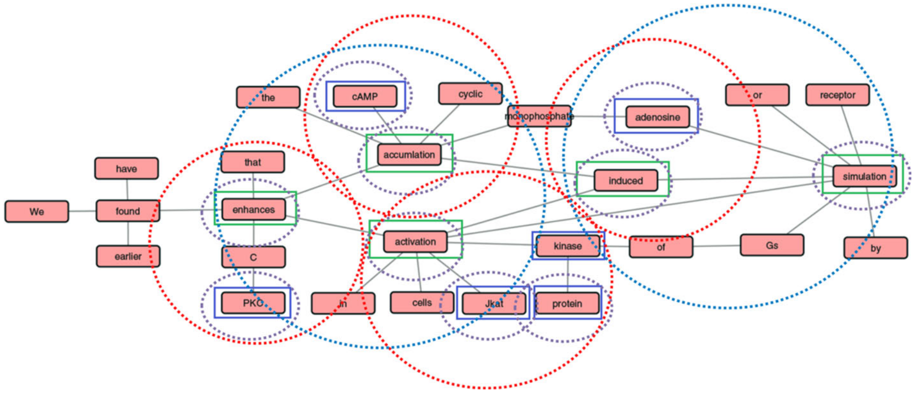


Fig. 7 Extraction of complex events by identification of association among subgraphs in MLG kernel. (The rectangle shape green represents trigger words and blue represents proteins. The dotted circles in various colors violet (Level 1), red (Level 2), blue (Level 3) represents each level of subgraph mining) (color figure online)

Example 1: PMID: 10330189

In response to activation of the Wnt signaling pathway, beta-catenin accumulates in the nucleus, where <**Keyword**> it </**Keyword**> cooperates with LEF/TCF (for lymphoid enhancer factor and T-cell factor) transcription factors to activate gene expression.

In the above example 1 the pronoun *"it"* denotes the protein *beta-catenin,* and it participated in the events *"gene expression"* and *"transcription"*.

Example 2: PMID: 10087185

Induction of NFkappaB is a **<Keyword>**highly**</Keyword>** regulated process requiring Phosphorylation.

In the above example 2, the keyword *"highly"* denoted the certainty of the event *"Phosphorylation"*.

The event-specific argument structure based syntactic rules were applied after contextual patterns and semantic role labeling to detect arguments. The event-specific argument structures-based rules acted as post-processing and improved the performance of the argument detection phase.

Even though our system performs well, it exhibits some limitations. The major source of errors that occurred in the argument detection phase is concerned with events containing multiple arguments. If the event contains more than three arguments, those types were difficult to extract. For example (PMID: 1313226), in the sentence "*Leukotriene B4 stimulates c-fos and c-jun gene transcription and AP-1 binding activity in human monocytes*". The event *regulation* contains more arguments and simultaneously it consisted of other events also as an argument.

4 Conclusions and future enhancements

In this paper, we deployed a hybrid system by combining methodologies such as the ensemble feature, graph-based kernels along with lexico-syntactic patterns to extract biomedical events from the literature. Our Multiscale Laplacian Graph (MLG) kernel-based approach can detect the topological relationships between events nodes in multiple scales and identifies the associations among the subgraphs for complex events. To the best of our knowledge, we are the first ones to introduce the MLG kernel for event extraction task. Since features play a crucial role in supervised machine learning, especially in event extraction a wide variety of features represented broadly as token-based, sentence based, parsing and domain-specific features to generate a feature-based kernel. Finally, we combined both ensemble kernels to generate a robust event classifier. In addition, in the argument detection phase we employed lexico-syntactic based semantic role labeling and contextual pattern-based rule engine to extract the event arguments. We incorporated contextual patterns, semantic role labeling, and event-specific argument structure to detect the domain-specific ontology-based concepts and relationships effectively. In the future, we plan to employ the automatic feature extraction approaches, advanced universal dependencies and different coefficient pair for kernel ensembling to extract the events from the literature, and we will apply this system in various biological relation extraction approaches such as Chemical Induced Disease (CID), Disease-Drug Interactions (DDIs) and Protein–Protein Interactions (PPIs).

Acknowledgements There is no separate funding received for this research work.

Data availability The link of source code and generated models is available at: http://biominingbu.org/bioevent_extraction/.

Compliance with ethical standards

Ethical approval Datasets used in the current work are all from BioNLP-shared tasks and GENIA, which are freely available for research work with suitable citations. Implementations of kernel approaches and Natural language processing methods used in the current work are all available as open source software with suitable citations.

Conflict of interest The authors declare that there are no conflicts of interest in this work.

References

1. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS (2015) Recent advances and emerging applications in text and data mining for biomedical discovery. Brief Bioinform 17(1):33–42
2. Cohen AM, Hersh WR (2005) A survey of current work in biomedical text mining. Brief Bioinform 6:57–71
3. Jesús Naveja J, Dueñas-González A, Medina-Franco JL (2016) Drug repurposing for epigenetic targets guided by computational methods. In: Medina-Franco José L (ed) Epi-informatics discovery and development of small molecule epigenetic drugs and probes. Academic Press, Cambridge, pp 327–357
4. Henry S, McInnes BT (2017) Literature based discovery: models, methods, and trends. J Biomed Inform 74:20–32
5. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel HP (2008) Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics 9(1):207
6. Murugesan G, Abdulkadhar S, Natarajan J (2017) Distributed smoothed tree kernel for protein–protein interaction extraction from the biomedical literature. PLoS ONE 12(11):e0187379
7. Bhasuran B, Natarajan J (2018) Automatic extraction of gene–disease associations from literature using joint ensemble learning. PLoS ONE 13(7):e0200699
8. Panyam NC, Verspoor K, Cohn T, Ramamohanarao K (2018) Exploiting graph kernels for high performance biomedical relation extraction. J Biomed Semantics 9(1):7
9. Zhou H, Ning S, Yang Y, Liu Z, Lang C, Lin Y (2018) Chemical-induced disease relation extraction with dependency information and prior knowledge. J Biomed Inform 84:171–178
10. Rios A, Kavuluru R, Lu Z (2018) Generalizing biomedical relation classification with neural adversarial domain adaptation. Bioinformatics 26(1):9

11. Vanegas JA, Matos S, Gonzalez F, Oliveira JL (2015) An overview of biomolecular event extraction from scientific documents. Comput Math Methods Med 015:571381
12. Ananiadou S, Pyysalo S, Tsujii JI, Kell DB (2010) Event extraction for systems biology by text mining the literature. Trends Biotechnol 28(7):381–390
13. Patumcharoenpol P, Doungpan N, Meechai A, Shen B, Chan JH, Vongsangnak W (2016) An integrated text-mining framework for metabolic interaction network reconstruction. PeerJ 4:e1811
14. Nawaz R, Thompson P, Ananiadou S (2013) Negated bio-events: analysis and identification. BMC Bioinformatics 14(1):14
15. Wang X, McKendrick I, Barrett I, Dix I, French T, Tsujii JI, Ananiadou S (2011) Automatic extraction of angiogenesis bioprocess from text. Bioinformatics 27(19):2730–2737
16. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J (2009) Overview of BioNLP'09 shared task on event extraction. In: Proceedings of BioNLP'09 shared task workshop, pp 1–9
17. Kim JD, Wang Y, Takagi T, Yonezawa A (2011) Overview of Genia event task in BioNLP shared task 2011. In: Proceedings of BioNLP shared task 2011 workshop, pp 7–15
18. Nedellec C, Bossy R, Kim JD, Kim JJ, Ohta T, Pyysalo S, Zweigenbaum P (2013) Overview of BioNLP shared task 2013. In: Proceedings of BioNLP shared task 2013 workshop, pp 1–7
19. Delėger L, Bossy R, Chaix E, Ba M, Ferrė A, Bessieres P, Nėdellec C (2016) Overview of the bacteria biotope task at bionlp shared task 2016. In: Proceedings of the 4th BioNLP shared task workshop 2016, pp 12–22
20. Kim JD, Ohta T, Tateisi Y, Tsujii JI (2003) GENIA corpus—a semantically annotated corpus for biotextmining. Bioinformatics 19:180–182
21. Thompson P, Nawaz R, McNaught J, Ananiadou S (2011) Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics 12(1):393
22. Zerva C, Batista-Navarro R, Day P, Ananiadou S (2017) Using uncertainty to link and rank evidence from biomedical literature for model curation. Bioinformatics 33(23):3784–3792
23. Le Minh Q, Truong SN, Bao QH. A pattern approach for biomedical event annotation. In: Proceedings of the BioNLP shared task 2011 workshop, pp 149–150
24. Kilicoglu H, Bergler S (2009) Syntactic dependency-based heuristics for biological event extraction. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task, pp 119–127
25. Liu X, Bordes A, Grandvalet Y (2013) Biomedical event extraction by multi-class classification of pairs of text entities. In: BioNLP shared task 2013 workshop, pp 45–49
26. Zhou D, He Y (2011) Biomedical events extraction using the hidden vector state model. Artif Intell Med 53(3):205–213
27. Li C, Liakata M, Rebholz-Schuhmann D (2013) Biological network extraction from scientific literature: state of the art and challenges. Brief Bioinform 15(5):856–877
28. Zhou D, Zhong D, He Y (2014) Event trigger identification for biomedical events extraction using domain knowledge. Bioinformatics 30(11):1587–1594
29. Lamurias A, Rodrigues MJ, Clarke LA, Couto FM (2016) Extraction of regulatory events using kernel-based classifiers and distant supervision. In: Proceedings of the 4th BioNLP shared task workshop, pp 88–92
30. Wang A, Wang J, Lin H, Zhang J, Yang Z, Xu K (2017) A multiple distributed representation method based on neural network for biomedical event extraction. BMC Med Inform Decis Mak 17(3):171
31. He X, Li L, Liu Y, Yu X, Meng J (2017) A two-stage biomedical event triggers detection method integrating feature selection and word embeddings. In: IEEE/ACM transactions on computational biology and bioinformatics
32. Jiang N, Rong W, Nie Y, Shen YK, Xiong Z (2017) Biological event trigger identification with noise contrastive estimation. IEEE/ACM Trans Comput Biol Bioinform 15:1549–1559
33. Bjorne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T (2009) Extracting complex biological events with rich graph-based feature sets. In: Proceedings of the workshop on current trends in biomedical natural language processing: shared task, pp 10–18
34. Bjorne J, Salakoski T (2013) TEES 2.1: automated annotation scheme learning in the BioNLP 2013 shared task. In: Proceedings of the BioNLP shared task 2013 workshop,pp 16–25
35. Hakala K, Van Landeghem S, Salakoski T, Van de Peer Y, Ginter F (2013) EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In: Proceedings of the BioNLP shared task 2013 workshop, pp 26–34
36. Riedel S, McClosky D, Surdeanu M, McCallum A, Manning CD (2011) Model combination for event extraction in BioNLP 2011. In: Proceedings of the BioNLP shared task 2011 workshop, pp 51–55
37. Lever J, Jones SJ (2016) VERSE: event and relation extraction in the BioNLP 2016 shared task. In: Proceedings of the 4th BioNLP shared task workshop, pp 42–49

38. Bjorne J, Salakoski T (2015) TEES 2.2: biomedical event extraction for diverse corpora. BMC Bioinform 16(16):4
39. Liu H, Komandur R, Verspoor K (2011) From graphs to events: a subgraph matching approach for information extraction from biomedical text. In: Proceedings of the BioNLP shared task 2011 workshop, pp 164–172
40. Liu H, Hunter L, Kešelj V, Verspoor K (2013) Approximate subgraph matching-based literature mining for biomedical events and relations. PLoS ONE 8(4):e60954
41. Liu H, Verspoor K, Comeau DC, MacKinlay AD, Wilbur WJ (2015) Optimizing graph-based patterns to extract biomedical events from the literature. BMC Bioinform 16(16):S2
42. Luo Y, Uzuner Ö, Szolovits P (2016) Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations. Brief Bioinform 18(1):160–178
43. Luo Y, Sohani AR, Hochberg EP, Szolovits P (2014) Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc 21(5):824–832
44. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P (2015) Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 22(5):1009–1019
45. Luo Y, Uzuner O (2014) Semi-supervised learning to identify UMLS semantic relations. In: AMIA summits on translational science proceedings, p 67
46. Zhang Y, Lin H, Yang Z, Wang J, Li Y (2013) Biomolecular event trigger detection using neighborhood hash features. J Theor Biol 7(318):22–28
47. Roberts K, Rink B, Harabagiu S (2010) Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task. In: Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data, i2b2 2010, Boston, MA, USA
48. Bùi QC (2012) Relation extraction methods for biomedical literature
49. Quirk C, Choudhury P, Gamon M, Vanderwende L (2011) Msr-nlp entry in bionlp shared task 2011. In: Proceedings of the BioNLP shared task 2011 workshop, pp 155–163
50. Dongliang X, Jingchang P, Bailing W (2017) Multiple kernels learning-based biological entity relationship extraction method. J Biomed Semant 8(1):38
51. Nikolentzos G, Siglidis G, Vazirgiannis M (2019) Graph Kernels: A Survey. arXiv preprint arXiv:1904.12218
52. Panyam NC, Verspoor K, Cohn T, Ramamohanarao K (2018) Exploiting graph kernels for high performance biomedical relation extraction. J Biomed Semant 9(1):7
53. Kondor R, Pan H (2016) The multiscale Laplacian graph kernel. In: Advances in neural information processing systems, pp 2990–2998
54. McClosky D, Surdeanu M, Manning CD (2011) Event extraction as dependency parsing. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, Vol 1, pp 1626–1635
55. Riedel S, McCallum A (2011) Robust biomedical event extraction with dual decomposition and minimal domain adaptation. In: Proceedings of the BioNLP shared task 2011 workshop 2011, pp 46–50
56. Munkhdalai T, Namsrai OE, Ryu KH (2015) Self-training in significance space of support vectors for imbalanced biomedical event data. BMC Bioinform 16(7):S6
57. Li L, Liu S, Qin M, Wang Y, Huang D (2016) Extracting biomedical event with dual decomposition integrating word embeddings. IEEE/ACM Trans Comput Biol Bioinform 13(4):669–677
58. Wang Y, Wang J, Lin H, Tang X, Zhang S, Li L (2018) Bidirectional long short-term memory with CRF for detecting biomedical event trigger in FastText semantic space. BMC Bioinform 19(20):507
59. Baldridge J (2005) The OpenNLP project. https://opennlp.apache.org/index.html. Accessed March 2015)
60. Liu H, Christiansen T, Baumgartner WA, Verspoor K (2012) BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. J Biomed Semant 3(1):3
61. Pado S, Lapata M (2007) Dependency-based construction of semantic space models. Comput Linguist 33(2):161–199
62. De Marneffe MC, MacCartney B, Manning CD (2006) Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC, pp 449–454
63. Sagae K, Tsujii JI (2007) Dependency parsing and domain adaptation with LR models and parser ensembles. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLPCoNLL)
64. Bhasuran B, Murugesan G, Abdulkadhar S, Natarajan J (2016) Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. J Biomed Inform 31(64):1–9
65. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, Lim S, Choi D, Kim S, Tan AC, Kang J (2016) BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PLoS ONE 11(10):e0164680

66. Murugesan G, Abdulkadhar S, Bhasuran B, Natarajan J (2017) BCC-NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition. EURASIP J Bioinform Syst Biol 2017(1):7
67. Claesen M, De Smet F, Suykens JA, De Moor B (2014) EnsembleSVM: a library for ensemble learning using support vector machines. J Mach Learn Res 15(1):141–145
68. Bjorne J, Salakoski T. Generalizing biomedical event extraction. In: Proceedings of the BioNLP shared task 2011 workshop, pp 183–191
69. Li Q, Ji H, Huang L (2013) Joint event extraction via structured prediction with global features. In: ACL, vol 1, pp 73–82
70. Campos D, Bui QC, Matos S, Oliveira JL (2014) TrigNER: automatically optimized biomedical event trigger recognition on scientific documents. Source Code Biol Med 9(1):1
71. Campos D, Matos S, Oliveira JL (2013) Gimli: open source and high-performance biomedical name recognition. BMC Bioinform 14(1):54
72. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007) Uniprotkb/swiss-prot. In: Edwards D (ed) Plant bioinformatics. Humana Press, Totowa, pp 89–112
73. Dunning T (2012) Finding structure in text, genome and other symbolic sequences. arXiv preprint arXiv:1207.1847
74. Naughton M, Stokes N, Carthy J (2008) Investigating statistical techniques for sentence-level event classification. In: Proceedings of the 22nd international conference on computational linguistics, vol 1. Association for Computational Linguistics, pp 617–624
75. Kondor R, Jebara T (2003) A kernel between sets of vectors. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 361–368
76. Chen Y, Hou P, Manderick B (2014) An ensemble self-training protein interaction article classifier. Bio-Med Mater Eng 24(1):1323–1332
77. Abdulkadhar S, Murugesan G, Natarajan J (2017) Classifying protein–protein interaction articles from biomedical literature using many relevant features and context-free grammar. J King Saud Univ Comput Inf Sci 32:553–560
78. Li L, Guo R, Jiang Z, Huang D (2015) An approach to improve kernel-based protein–protein interaction extraction by learning from large-scale network data. Methods 15(83):44–50
79. Hung SH, Lin CH, Hong JS (2010) Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labeling. Expert Syst Appl 37(1):341–347
80. Hearst MA (1992) Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on computational linguistics, vol 2. Association for Computational Linguistics, pp 539–545
81. Miwa M, Sætre R, Kim JD, Tsujii JI (2010) Event extraction with complex event classification using rich features. J Bioinform Comput Biol 8(01):131–146
82. Bui QC, Campos D, Van Mulligen E, Kors J (2013) A fast rule-based approach for biomedical event extraction. In: Proceedings of the BioNLP shared task 2013 workshop, pp 104–108
83. Björne J, Salakoski T (2018) Biomedical event extraction using convolutional neural networks and dependency parsing. In: Proceedings of the BioNLP 2018 workshop, pp 98–108
84. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S (2012) Extracting semantically enriched events from biomedical literature. BMC Bioinform 13(1):108

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sabenabanu Abdulkadhar is a PhD research scholar at Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore. She has completed her Master's in Computer Applications from Anna University, Coimbatore. After that, she joined as a JRF under DBT project in Data Mining and Text Mining Laboratory, Department of Bioinformatics, she is pursuing her PhD in biomedical text mining. Her research work focuses on Machine learning, Data mining, Text mining, Text classification and Bio statistics.

Balu Bhasuran is a PhD research scholar at DRDO-BU Center for Life Sciences, Bharathiar University campus, Coimbatore, TamilNadu, India. He has completed Masters in Computer Applications from Mahatma Gandhi University, Kerala. After that, he joined as a JRF at DRDO-BU Center for Life Sciences and currently pursuing his PhD in biomedical text mining. His research interests are in the areas of big data analytics, text mining, network analysis, knowledge discovery, machine learning.

Jeyakumar Natarajan is currently working as Professor and Head at Department of Bioinformatics, Bharathiar University, Coimbatore, India. He completed his PhD in Bioinformatics from University of Ulster, Belfast, United Kingdom and spent one year as a visiting predoctoral fellow at Northwestern Medical School, Northwestern University, Chicago, USA. He established the Data Mining and Text Mining Research Group at Department of Bioinformatics, Bharathiar University and mainly working on data mining, text mining and machine learning methods for high-throughput biomedical data analysis and interpretation.