Biomolecular Event Extraction using Natural Language Processing

Review Paper

Manish Bali
Department of Computer Science and Engineering,
Presidency University, Bengaluru,
Karnataka, India
balimanish0@gmail.com

S. P. Anandaraj
Department of Computer Science and Engineering,
Presidency University, Bengaluru
Karnataka, India
anandsofttech@gmail.com

Abstract *– Biomedical research and discoveries are communicated through scholarly publications and this literature is voluminous, rich in scientific text and growing exponentially by the day. Biomedical journals publish nearly three thousand research articles daily, making literature search a challenging proposition for researchers. Biomolecular events involve genes, proteins, metabolites, and enzymes that provide invaluable insights into biological processes and explain the physiological functional mechanisms. Text mining (TM) or extraction of such events automatically from big data is the only quick and viable solution to gather any useful information. Such events extracted from biological literature have a broad range of applications like database curation, ontology construction, semantic web search and interactive systems. However, automatic extraction has its challenges on account of ambiguity and the diverse nature of natural language and associated linguistic occurrences like speculations, negations etc., which commonly exist in biomedical texts and lead to erroneous elucidation. In the last decade, many strategies have been proposed in this field, using different paradigms like Biomedical natural language processing (BioNLP), machine learning and deep learning. Also, new parallel computing architectures like graphical processing units (GPU) have emerged as possible candidates to accelerate the event extraction pipeline. This paper reviews and provides a summarization of the key approaches in complex biomolecular big data event extraction tasks and recommends a balanced architecture in terms of accuracy, speed, computational cost, and memory usage towards developing a robust GPU-accelerated BioNLP system.*

Keywords*: Bimolecular event extraction, natural language processing, text mining, machine learning, BioNLP shared task*

1. INTRODUCTION

Medical literature is a vast repository for knowledge sharing that happens in the biomedical domain. With major advances in computational biology and allied scientific research, there is an explosion in the number of publications in this area [1]. Every day approximately three thousand research articles are getting published in biomedical journals. Considering just one database, say MEDLINE, there are 23,000,000 references with 40,000-50,000 getting added every day. For any researcher, this poses an enormous challenge to locate, manage and choose suitable literature in their domain. Thus, the automated mechanism to extract structured content (explicit knowledge) from unstructured text (implicit knowledge) as shown in (Fig. 1) is the need of the hour [2]. The information deluge is posing new challenges as bio-databases, vocabularies and bio-ontologies encode only a small fraction of information. Curators are struggling to process scientific literature. The discovery of facts and events is crucial for gaining insights into biosciences, hence the need for text event mining. Artificial Intelligence and especially Machine Learning techniques such as NLP and TM tools have gained significant importance to curate large biological databases. This has led to the development of many new applications and search engines addressing various domains for mining databases [3-5].

Previous research has focused on the basic extraction of entities and identifying their links in reference knowledge bases [6, 7]. Few existing techniques though provide acceptable performance [8] for many applications. Off-late interest has arisen in biomedical entities, for example, drug-protein, drug-drug, and protein-protein interactions [10,11] which have emerged as the most important entities due to several similar databases and their usage in systems biology.

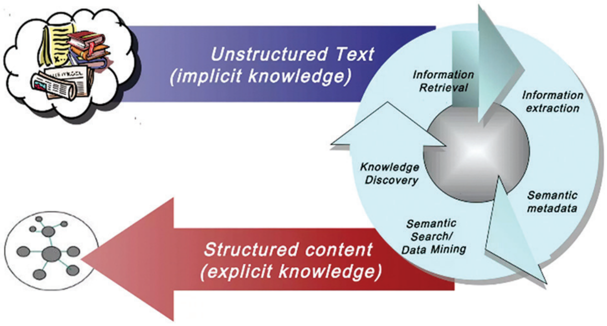

Fig. 1. Solving the biodata deluge using text mining (TM) [9]

NLP-based data-channeling process and their potential applications are many [12,13] as illustrated in (Fig. 2). Initially in the pipeline, anyone working in the biomedical NLP domain uses Information retrieval (IR) techniques which include tasks such as classification and retrieval of documents to select relevant articles. This process helps to reduce unwanted text or documents to only those that are of interest to the researcher. Next in the pipeline, Information extraction (IE) techniques are used to primarily identify any named entity and/or extract any event that is of researcher's interest and provides useful information. Some examples of information that are typically identified, mined, or extracted from biomedical literature include an entity interacting with an entity like a drug interacting with a drug or a protein interacting with a protein. It also includes a relationship between an entity with another entity like between a protein and a residue or a gene or any temporal relationship in addition to any other bio-entities that are part of the event. Hence, this process automates and simplifies the task of offering only useful textual data required to the researcher by eliminating cumbersome manual searching efforts [14,15]. These mined events from vast literature are important and have many real-world applications like database curation, constructing ontology, semantic web search, interactive systems etc.

Named entity recognition (NER), is a subset of the event extraction task that involves identifying and detection of references to entities like genes, and proteins [16]. It is also a research area of interest which has gained a lot of traction over the last decade.

This is because there is still a large gap of >10% in the F1- score using the best ML algorithms for biomedical NER vs. those used for any general-purpose NER. Hence, researchers are exploring various methods to narrow the gap using better pre-processing and feature extraction techniques. Most approaches for effectively identifying named entities (NEs) in biomedical literature fall under three categories, namely heuristic rule-based, dictionary-based, and statistical machine learning-based approaches. But the results to date have been far from satisfactory which suggests that there is still no robust, generalized implementation of any NER system nor any algorithm which can be singled out that can provide higher performance.

This paper is divided into the following sections: Section 2 explains a biomolecular event extraction task, and Section 3 discusses and compares the performance of the existing systems. Section 4 discusses new architectures and suggests two novel approaches for developing a robust BioNLP model. We conclude with a note summarizing the approaches, existing challenges, and future research directions in Section 5.

2. BIOMOLECULAR EVENT EXTRACTION

Identifying and isolating semantic relations is the primary task when it comes to biomedical text mining. Here, information is extracted from a vast volume of document sets or big data like scientific literature or patient records [17-19]. The information contains, apart from other things, statements of interactions between NEs, like the effect of drugs on a patient, cellular protein movements etc. Relation extraction and event extraction are the two ways for getting such data. Relations can be typed, directed, or pairwise links type between defined named entities.

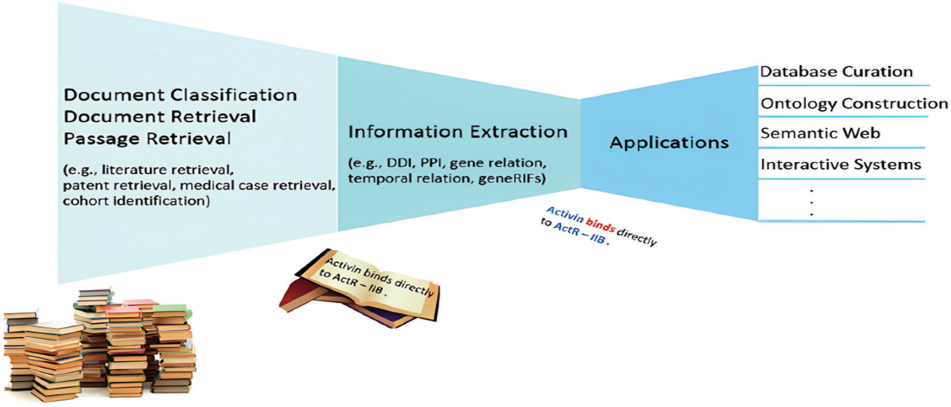
Document Classification
Document Retrieval
Passage Retrieval
(e.g., literature retrieval, patent retrieval, medical case retrieval, cohort identification)
Information Extraction
(e.g., DDI, PPI, gene relation, temporal relation, geneRIFs)
Applications
Database Curation
Ontology Construction
Semantic Web
Interactive Systems
.
.
.
Activin binds directly to ActR – IIB .
Activin binds directly to ActR – IIB .

Fig. 2. NLP-based data-channeling process and applications [20]

Event extraction is another form of relation extraction in which events can combine two or more entities, they have a trigger word which is usually a verb and they also sometimes act as arguments for other events [21]. Though events are efficient in capturing the semantics of text more precisely extracting them is an extremely complicated task. Table 1 explains the event extraction process. Post-pre-processing and feature extraction, first, the NEs (e.g., TGF-beta) are identified. The next stage is to detect trigger words and labels via annotation of the phrases and finally the process reconstructs the events if the edges are clear or not overlapping.

A typical biomedical event extraction task flow-chart is shown in (Fig. 3). The first two steps are pre-processing and feature extraction followed by named entity recognition [22]. Trigger detection (which identifies event triggers and types) and edge detection (links event triggers with arguments) are the two sub-steps of the main event detection step. Some researchers combine these two steps to reduce cascading errors which helps in improving performance. Post-processing is the final stage that helps in refining the final event structure outcome [23,24].

All the steps are explained in the following sections outlining various tools and approaches used.

Table 1. A biomolecular event extraction task workflow

Phrase: "TGF-beta mediates RUNX induction and FOXP3 is efficiently up-regulated by RUNX1 and RUNX3" **Entities: "TGF-beta, RU NX, FOXP3, RUNX1 and RUNX3 (proteins)"**		
Step 1	Entity recognition	Entities: TGF-beta, RUNX, FOXP3, RUNX1 and RUNX3 (proteins)
Step 2	Trigger detection	Trigger 1: induction (gene expression)
		Trigger 2: mediates (positive regulation)
		Trigger 3: up-regulated (positive regulation)
Step 3	Edge detection	Edge 1: induction (gene expression); theme: RUNX(protein)
		Edge 2: mediates (positive regulation); cause: RUNX (protein); theme: above gene expression event
		Edge 3: up-regulated (positive regulation); cause: RUNX1 and RUNX3; theme: FOXP3 (protein)
Step 4	Reconstruct event	Event 1: induction (gene expression); theme: RUNX (protein)
		Event 2: mediates (positive regulation); cause: RUNX (protein); theme: above gene expression event
		Event 3: up-regulated (positive regulation); cause: RUNX1; theme: FOXP3 (protein)
		Event 4: up-regulated (positive regulation); cause: RUNX3; theme: FOXP3 (protein)

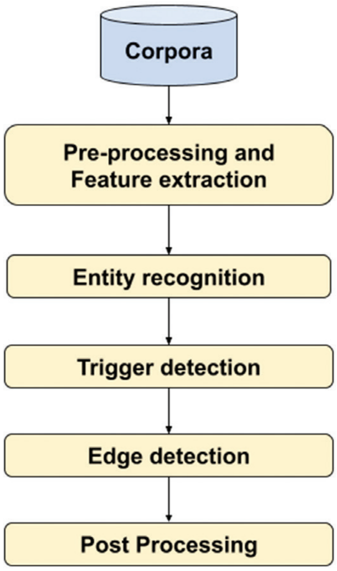

Fig. 3. Flow-chart of a typical biomolecular event extraction task

2.1. CORPORA

There are many corpora available for biomolecular event extraction tasks. Some of the most popular ones used are enumerated below:

- GENIA Event dataset [25]: This is made available by the BioNLP shared task [26] organizers openly to all researchers. This consists of human-curated complex event events. It has 1000 Medline paper abstracts, which in turn have 9372 sentences and 36114 events have been identified from it.
- BioInfer Dataset [27]: It is a publicly available dataset consisting of manually annotated corpus and other resources for extraction of information. It has 1,100 sentences from biomedical research abstracts.
- PPI dataset [28]: A Protein-protein interaction corpus complements available training data and is not as elaborate as the event corpora. LLL, AIMed and BioCreative are the most relevant PPI corpora used.

2.2. PRE-PROCESSING AND FEATURE EXTRACTION

Datasets are getting more complex by the day and it may be required to work with datasets containing hundreds of features. If the number of features equals or becomes more than the number of observations stored in a dataset, it could lead to overfitting. Hence pre-processing and applying feature extraction techniques are necessary. Pre-processing, a mandatory step in any text mining pipeline involves reading the data from its original format to an internal representation. It involves a set of common NLP tasks, from sentence segmentation and tokenization to part-of-speech tagging, chunking, and linguistic parsing. In addition to these, the biomolecular event extraction task also involves removing co-references, sentence simplification etc. to improve accuracy [29,30]. The most commonly used pre-processing frameworks are NLTK (http://www.nltk.org/), Stanford CoreNLP (http://nlp.stanford.edu/software/corenlp.shtml) and Apache OpenNLP (https://opennlp.apache.org/).

Table 2. Commonly used features in the event detection phase

Feature Groups	Features	Trigger recognition	Edge detection
Token-based	Parts-of-speech (POS)	Yes	Yes
	Lemma	Yes	No
	Orthographic	Yes	No
	n-grams	Yes	No
	Word shape	Yes	No
	Prefixes/suffixes	Yes	No
Contextual features	Number of entities	Yes	No
	BoW counts	Yes	No
	Windows or conjunctions of features	Yes	No
Dependency-based features	Number and type of dependency edges	Yes	No
	Words, lemmas, and POS tags in the dependency path	Yes	Yes
	N-grams in the dependency path	Yes	Yes
External features	WordNet lemmas	No	No
	Trigger lexicon	No	No
	Entity lexicon	No	No

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones, which are then discarded. Feature extraction techniques provide below advantages:

- Accuracy improvement [31-34]
- Overfitting risk reduction
- Speed up in training
- Improved data visualization
- Increase in explainability of a model

There are many feature extraction techniques used. The commonly used features in the main event detection stage are shared in Table 2. They can be divided into four feature groups namely token-based, contextual, dependency and external resources which are explained below.

- Token-based features: They capture a specific feature for every token like part-of-speech, lemma, prefix, suffix etc. In part-of-speech, each word in a sentence is tagged with its POS like noun, verb, adjective etc. However, its main drawback is ambiguity. In the biomedical domain, many frequently used words have multiple meanings hence multiple POS. Lemmatization breaks a word down to its root meaning to identify similarities. But it is a slow and time-consuming process as it involves performing morphological analysis and deriving the meaning of words from a dictionary. Orthographic rules are general rules used when breaking a word into its stem and modifiers. However, many languages have various levels of orthographic depth and orthographies that are highly irregular, and difficult, and where sounds cannot be predicted from the spelling. n-grams refer to a sequence of N words or characters. More is not necessarily better as in some cases, having too many features will result in a less optimal model. Base forms of the input words, word prefixes and suffix features, and basic word form/shape features have also been used for trigger recognition with varying degrees of success.
- Contextual features: They offer knowledge about the sentence or its neighborhood where the token exists. The same word or phrase can also have different meanings according to the context of a sentence or many words. The number of tokens in the sentence, no. of entities or the bag-of-word count are some of the contextual features extracted from sentences. A bag of words is a representation of text that describes the occurrence of words within a document. It just keeps track of word counts and disregards the grammatical details and the word order. It is called a 'bag' of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. A drawback of the BoW model is that it ignores the location information of the word. Also, the model does not respect the semantics of the word. The range of vocabulary is a big issue faced by the bag-of-words model.
- Dependency-based features: They provide the Grammar relation between two words and can be extracted from a dependency relationship graph of a sentence. Commonly used features include the number and type of dependency edges between two tokens and the words, lemmas, POS and n-gram characteristics in the dependency path. The drawbacks of dependency parsing are first, semantic actions cannot be performed while making a prediction. The actions must be delayed until the prediction is known to be a part of a successful parse. Secondly, precise error reporting is not possible. A mismatch merely triggers backtracking.
- External features: Here external knowledge is augmented from sources like WordNet, lexicons both trigger and entity from the dictionary or other sources. This helps to improve accuracy in cases where the no. of words is long, has many meanings etc. In linguistics, the canonical form or morphological form of a word is called a lemma. To find a synonym as well as an antonym of a word, one can look up lemmas in WordNet. The drawbacks are that the vocabulary it contains is broad and thus

ambiguous. Lexicons are collections of domain-specific key phrases (also known as entities) that can be attached to a flow. A Lexicon can be seen as a dictionary, which allows the virtual agent to 'understand' specific words. The drawback of a lexicon-based feature extraction model is scalability.

2.3. NAMED ENTITY RECOGNITION

NER detects references to entities like genes, proteins, and chemical compounds [35,36]. It is a challenging task in the biomedical domain as new entities keep getting added and there is a lack of a complete dictionary. Also, it may so happen that a word may refer to two different entities based on their context. It is also seen that many biological NEs have different spelling forms, have abbreviations or are exceptionally long which come in the way of classifying them or identifying their boundaries correctly. They may also be nested in other entities, which require more effort to identify such NEs. There are a few NER systems proposed, but the best results are in the 85% F1-score range [37]. Some of the most popular NER tools used are summarized in Table 3.

Table 3. Most popular NER tools used

Gene/ Proteins	Chemical compounds	Gene protein and disorders
Gimli	SCAI	BioEnEx
NERSuite	ChemSpot	BANNER
AIIAGMT	Neji	-
GNAT	-	-
GeNo	-	-
ABNER	-	-

2.4. TRIGGER DETECTION

A lot of research is focused on this area as the subsequent steps' effectiveness is based on the information outcome from this step. It involves identification of event triggers and their types. As seen from Table 1, the trigger word 'induction' defines an event of type gene expression and the trigger word 'mediates' defines an event of type Positive Regulation. Complexity arises when sentences may contain many related events, Negative Regulation and the same trigger could also indicate diverse types of events, based on the context [38-40]. Table 4 summarizes the different methodologies adopted for Trigger detection which are broadly classified as rule-based based, dictionary-based and machine learning based. It is observed that rule-based methods return low recall rates as defining detailed rules requires a lot of effort and some of the rules are too hard to accommodate semantic paraphrases. Dictionary-based approaches contain a dictionary with trigger words and their corresponding classes (event types) to identify and enumerate event triggers. Researchers have also used a hybrid approach by combining rule-based and dictionary-based approaches [41,42].

Table 4. Different methodologies adopted for trigger detection

Rule-based	Dictionary-based	Machine Learning based				References
		SVM	CRF	VSM	MEMM	
√	√	×	×	×	×	[51]
×	√	×	√	×	×	[8]
×	×	√	×	×	×	[52]
×	√	×	×	×	×	[53]
√	√	×	×	×	×	[41]
√	√	√	×	×	×	[54]
√	×	×	×	×	×	[55]
√	×	√	√	×	×	[56]
×	×	√	×	√	×	[57]
×	×	×	√	×	√	[58]
×	×	×	√	×	×	[59]
×	×	×	√	×	×	[60]
×	×	×	√	×	×	[61]
×	×	×	√	×	×	[62]
×	×	×	√	×	×	[63]
×	×	×	√	×	×	[64]
×	×	×	√	×	×	[65]
×	×	×	×	√	×	[66]

Machine Learning based approaches like Support vector machine (SVM) and its variant kernels like linear, radial basis function, polynomial and convolution; Conditional random field (CRF); Value stream matching (VSM) and Maximum entropy Markov model (MEMM) have been most successful for trigger detection and offer higher recall rates [43-46].

2.5. EDGE DETECTION

It predicts arguments in an event which can be named entities, or another event represented as a different trigger word. It is also known as event theme construction. In Table 1, for the sentence, there are three edges identified. Again, rule, dictionary and machine learning-based methods have been suggested to address this task. Like trigger detection, there have been more efforts in using ML algorithms by treating the edge detection problem as a supervised multi-class classification problem. Also, many studies are based on hybrid approaches by using a combination of the above methods or using ensemble methodology [47,48] as shown in Table 5.

Table 5. Hybrid approaches used for trigger detection

Ensemble	Base Learners	CI-Optimized algorithms	F1-score (%)	References
Yes	1	Yes	55.64	[67]
Yes	3	Yes	57.58	[47]
Yes	5	No	66.34	[68]

Here, the thought process is not to choose the 'best' classifier always, as it may not be representative of all data. In this, a few classifiers are trained in a standalone mode. The output of all the classifiers is combined into an ensemble and the result is chosen according to some criteria. A problem arises on how to choose the best one, as more than one classifier may also meet the criteria of similar training accuracy. Methods of the combination include bagging, boosting, voting, stacked generalization, and cascading. This leads to improvement in prediction accuracy [49].

Researchers have also used optimization techniques to choose the right weights for voting using various nature-inspired Computational Intelligence (CI) algorithms [50]. Table 6 summarizes the various approaches used for Edge detection.

Table 6. Various approaches used for Edge detection

Rule-based	Dictionary-based	Machine Learning based			References
		SVM	CRF	HVS	
√	√	×	×	×	[54]
√	×	√	×	×	[60]
√	×	×	×	×	[8]
×	×	√	×	×	[53]
√	×	×	×	×	[41]
√	×	×	×	√	[51]
×	×	√	×	×	[58]
×	×	×	√	×	[8]
×	×	√	×	×	[59]
×	×	√	×	×	[60]
×	×	√	×	×	[62]
×	×	√	×	×	[63]
×	×	√	×	×	[69]

3. EXISTING SYSTEM

Many bio-text mining campaigns are running successfully for years. Some of the most popular ones are listed below. They address many aspects from NER to the biological phenomenon to text categorization

- KDDCup
- TREC-Genomics
- JNLPBA for NER
- BioCreative for extraction of NER, PPI, text categorization
- BioNLP Shared tasks

We discuss and compare the existing state-of-the-art systems and the approaches used in this section. For a similar comparison, the results achieved by various systems for event extraction on the BioNLP shared tasks are compared [70,71]. The BioNLP-ST uses the Genia event (GE) corpus with sub-tasks. It extracts events from both complete papers and abstracts. The subtasks are divided into three categories, (i) Core event extraction (GE), (ii) Event enrichment (GE 2) and (iii) Negation/Speculation detection (GE 3). Table 7 lists the shared tasks released by the committee in 2013 [72-74].

Table 8. Results obtained in the BioNLP-ST GE task by top systems (in %age)

System, Year, Reference	Simple	Event type binding	Regulation	Total
Univ. of Turku, 2009 [52]	70.21	44.41	40.11	52
Miwa, 2010 [53]	70.44	52.62	40.6	53.3
FAUST, 2011 [75]	72.85	51.05	46.97	57.5
EVEX [63]	76.59	42.88	38.41	51
TEES 2.1, 2013 [76]	76.82	43.32	38.05	50.7
BioSM [77]	76.11	49.76	35.8	50.7

Table 9. Results from BioNLP-ST GE 2 task by top systems (in %age)

System, Year, Reference	Site	Localization	Total
Univ. of Turku, 2009 [52]	71.43	36.59	44.5
FAUST, 2011 [75]	50	50	52.8
EVEX, 2013 [63]	50	50	31.2
TEES 2.1 [76]	50	50	32.5

Table 10. Results obtained in the BioNLP-ST GE 3 task by top systems (in %age)

System, Year, Reference	Document	Negation	Speculation	Total
Concord, 2009 [54]	Abs	23.13	25.27	24.17
Univ. of Turku, 2011 [55, 78]	Abs	30.4	25.64	28.08
EVEX, 2013 [63]	FP	27.04	23.92	25.22
TEES 2.1 [77]	FP	27.3	23.61	25.15

Table 11. Benchmark performance of training GloVe on CPU vs. GPU

Architecture	Specifications	Performance	Speed-up
CPU	i7-6800K, 8-cores	13.56 min/ epoch	-
GPU	NVidia GTX 1070	1.22 min/ epoch	11X

4. NEW ARCHITECTURES

Implementing CNNs and RNNs, though help in improving accuracy but is impacted on account of training time, computational cost and memory requirements. BERT (Bidirectional encoder representations from Transformers) which is an encoder stack of the

Transformer architecture has emerged as a strong contender for BioNLP tasks as it is pre-trained on a large but generic corpus. However, to make it suitable for BioNLP, it needs to be suitably trained on additional Biomedical corpus which is time-consuming. Hence, newer hardware architectures are required as an alternative to standard x86 processors. High-performance accelerators exist for many tasks today, that can be explored for NLP too. Current models focus only on accuracy and seldom on the issues concerning the above three impact criteria mentioned. NVidia Graphical processing unit (GPU), Intel Xilinx Field programmable gate array (FPGA) and Google's Tensor Processing Units (TPU) are some of the accelerator options that provide high-performance computing (HPC) at a fraction of the cost. We discuss the benefits of one such accelerator in this paper, the Nvidia GPU, compare it with a standard x86 processor and explain briefly how it can accelerate complex BioNLP tasks.

4.1. GPU ACCELERATED BIONLP

Many tasks in NLP can exploit the massive parallelism offered by the GPU. GPUs were initially used only for graphics or visualization tasks. Due to the massive number of lightweight cores (SMs) as shown in (Fig. 4), in early 2000 Nvidia introduced the CUDA parallel programming environment, which revolutionized high-performance computing. CUDA is C-like programming which allowed researchers to port their compute-hungry codes onto the GPU. Single-instruction multiple-thread (SIMT) execution model emerged and with the introduction of shared memory and a heterogeneous CPU-GPU architecture, all heavy-duty compute or mathematical tasks were off-loaded to the GPUs for processing thereby accelerating the entire code.

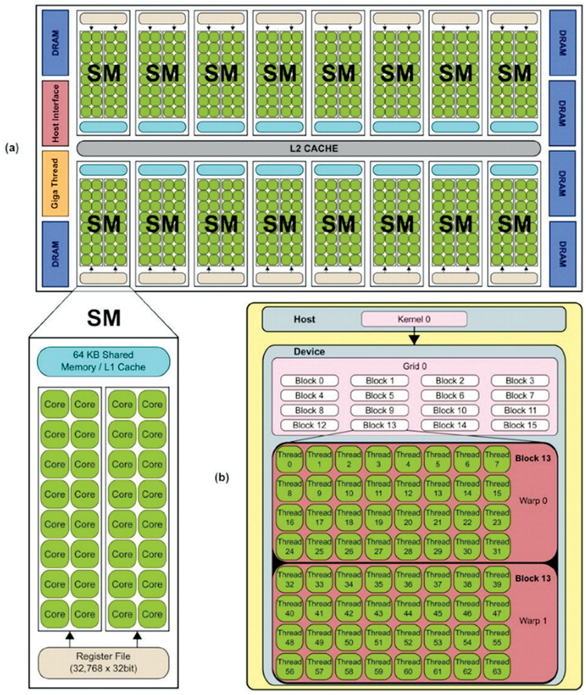

Fig. 4. The Nvidia GPU architecture [79]

Many GPU-accelerated NLP toolkits exist including GloVe (Global vectors for word representation), which is an extremely popular unsupervised learning algorithm for words. Especially for NLP, once the text is hashed, GPUs can offer accelerated results from the voluminous scientific literature much faster (large no. of words per millisecond) as compared to an x86 CPU. (Fig. 5) shows the difference between a CPU which has lesser ALUs compared to a GPU which explains this significant difference in computational power.

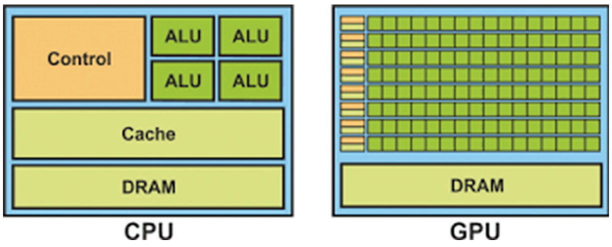

Fig. 5. Differences between a CPU and GPU [79]

The goal of a CPU is different from a GPU in that a CPU must be capable of processing everything, serial or parallel but the GPUs are grounds-up built for parallel tasks only, and NLP falls in this category. One of the drawbacks of using GPUs is the precision overhead. Precision is particularly important in operations like NLP, computing gradients etc. It is found that GPU speeds with double precision are 2-32x slower. Also, it consumes a lot of memory and power. The solution for this lies in using mixed precision for NLP tasks and lower-power gaming cards like the GTX series. Table 11 shows the benchmark performance of training GloVE on a CPU vs. a CPU+GPU system.

Very few implementations of CNN and RNN models have been tried on accelerators, which can help achieve up to 10 Tera operations (TOPs) per second on such semiconductor chips [80]. Each of these chips has its unique hardware architecture and programming environments to launch thousands of threads to parallelize the applications. But we still need large-scale HPC clusters with large numbers of nodes in a datacenter to achieve an inference efficiency of millions of words per second. Authors in [81] implemented a Dynamic multi-pooling convolutional neural network (DMCNN) that took 1.0 GOPs for processing 30 words in a single sentence. Normal CPUs are unable to keep pace with RNN computations on account of irregular computation and memory access. These accelerators though desired have some limitations as they are rated at higher power and provide lower performance per watt as the complexity of neural networks increases. Thus, achieving a balance of accuracy, computational cost and memory usage needs to be targeted and this is an area where not much research has been conducted.

4.1. RECOMMENDED ARCHITECTURES

Based on an extensive literature study and results of best-performing models, an ensemble model of ML classifiers and a domain-specific BERT pre-trained using GPU for reducing the training time on additional biomedical corpora (like PubMed and PubMed Central extracts) are the two key architectures recommended for NER, the most challenging sub-task in Biomolecular event extraction. We propose high-level reference architecture and methodology of an ensemble classification model and NeRBERT- a biomedical NER tagger in (Fig. 6) and (Fig. 7) respectively.

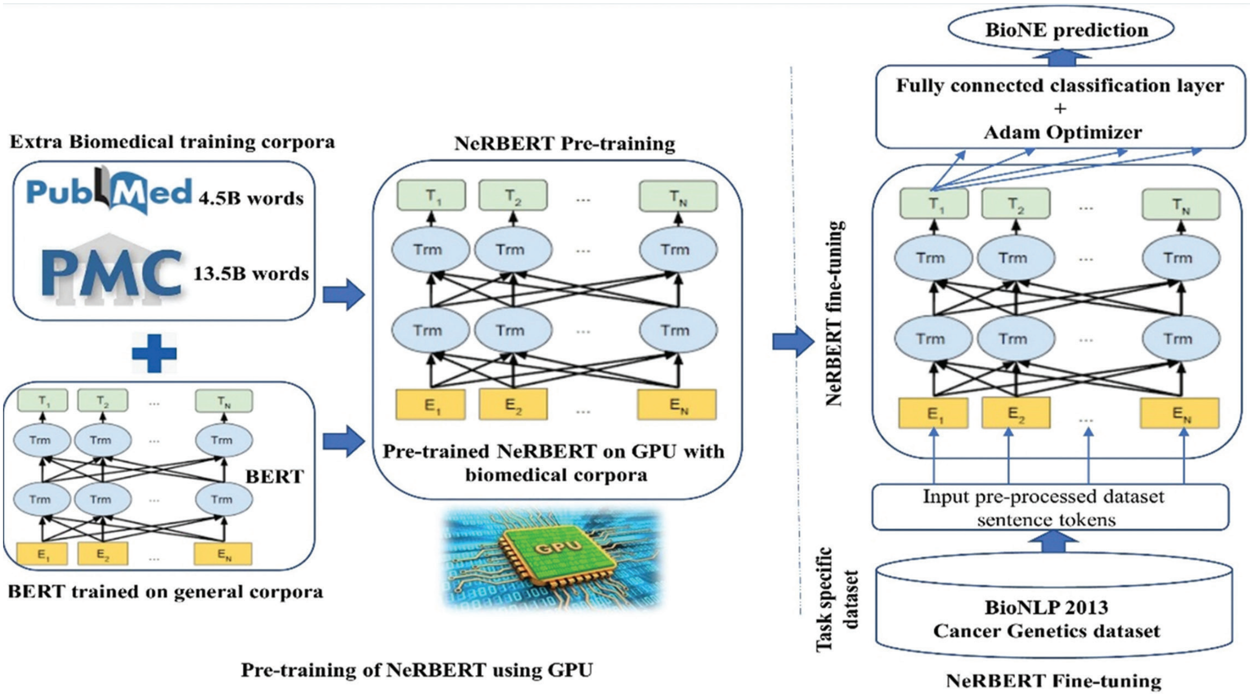
Extra Biomedical training corpora
PubMed 4.5B words
PMC 13.5B words
BERT
BERT trained on general corpora
NeRBERT Pre-training
T_1 T_2 ... T_N
Trm Trm ... Trm
E_1 E_2 ... E_N
Pre-trained NeRBERT on GPU with biomedical corpora
GPU
Pre-training of NeRBERT using GPU
NeRBERT fine-tuning
Task specific dataset
BioNE prediction
Fully connected classification layer + Adam Optimizer
Input pre-processed dataset sentence tokens
BioNLP 2013 Cancer Genetics dataset
NeRBERT Fine-tuning

Fig. 7. Recommended architecture of a pre-trained and fine-tuned NeRBERT model

5. CONCLUSION

The paper provides a systematic summarization of the latest research in the field of biomolecular event extraction. It discusses the various techniques used by researchers due to the challenging nature of the task, ambiguity, and heterogeneity in biomedical literature. The results summarized demonstrate that this is still a challenging proposition, despite the slow and steady improvements. Although performance results of 80% in the F1-score were obtained in the identification of simpler events, there is still minimal extraction of more complex events such as binding and regulatory events. Although major efforts have been made to recognize the events, the best performance achieved remains 30–40% lower than that for simple events. Many techniques have been proposed, which include simple parsing, pattern-matching, machine learning and deep learning methods. Researchers have adopted multi-stage approaches wherein the second stage fine-tunes the output from the first stage either by using some rules or techniques like ensemble classification which is one of the architectures recommended. Alternatively, BERT has also emerged as an architecture with great promise. We propose two reference architectures based on these in terms of accuracy, computational cost, and memory usage for developing a robust GPU-accelerated BioNLP model for biomolecular event extraction. The BERT model should be domain-specific and trained on additional biomedical corpora using a GPU to reduce training time. However, key challenges to mitigate still exist which involve extracting complex regulatory events, resolving cross-references, and defining negation and speculation. Also, machine learning approaches like transfer learning and other newer hardware architecture like FPGA and TPU have not been employed much which provide a lot of scopes to accelerate the event extraction pipeline. Despite these challenges, available research can still help in curating pipelines using text-mined data, constructing networks, ontologies, and knowledge bases.

6. REFERENCES:

[1] M. S. Simpson, D. Demner-Fushman, “Biomedical text mining: a survey of recent progress”, Proceedings of Mining Text Data, Springer, 2012, pp. 465-517.

[2] C. Li, M. Liakata, D. Rebholz-Schuhmann, “Biological network extraction from scientific literature: state of the art and challenges”, Briefings in Bioinformatics, Vol. 15, No. 5, 2014, pp. 856- 877.

[3] A. Manconi, E. Vargiu, G. Armano, L. Milanesi, “Literature retrieval and mining in bioinformatics: state of the art and challenges”, Advances in Bioinformatics, 2012, pp. 10-10.

[4] S. Ananiadou, P. Thompson, R. Nawaz, "Event-based text mining for biology and functional genomics", Briefings in Functional Genomics, Vol. 14, No. 3, 2015, pp. 213-230.

[5] L. Hirschman, G. A. P. C. Burns, M. Krallinger, "Text mining for the Biocuration workflow", Database: The Journal of Biological Databases and Curation, 2012.

[6] D. Campos, S. Matos, J. L. Oliveira, "Current methodologies for biomedical named entity recognition", John Wiley & Sons, 2013, pp. 839-868.

[7] C. N. Arighi, Z. Lu, M. Krallinger, "Overview of the Biocreative III workshop", BMC Bioinformatics, Vol. 12, 2011, Supplement 8, S1.

[8] D. Mackinlay, D. Martinez, T. Baldwin, "Biomedical event annotation with CRFs and precision grammars", Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, June 2009, pp. 77-85.

[9] E. Ghosh, H. Naja, H. Abdulrab, M. Khalil, "Ontology Learning Process as a Bottom-up Strategy for Building Domain-specific Ontology from Legal Texts", Proceedings of the 9th International Conference on Agents and Artificial Intelligence, 2017, pp. 473-480.

[10] I. Segura-Bedmar, P. Martínez, M. Herrero-Zazo, "SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts", Proceedings of the Seventh International Workshop on Semantic Evaluation, Atlanta, GA, USA, June 2013, pp 341-350.

[11] S. Ananiadou, S. Pyysalo, J. Tsujii, D. B. Kell, "Event extraction for systems biology by text mining the literature", Trends in Biotechnology, Vol. 28, No. 7, 2010, pp. 381-390.

[12] A. Prodromidis, P. K. Chan, P, S. J. Stolfo, "Meta-learning in distributed data mining systems: Issues and approaches", Advances in distributed and data mining, AAAI Press, 2016.

[13] P. Thompson, S. A. Iqbal, J. Mcnaught, S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction", BMC Bioinformatics, Vol. 10, 2009, pp. 349-349.

[14] D. Zhou, S. Jian, "Deep Knowledge Resources in Biomedical Name Recognition", Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, 2014.

[15] H. G. Lee, H. C. Cho, M. J. Kim, J. Y. Lee, G. Hong, H. C. Rim, "A multi-phase approach to biomedical event extraction", Proceedings of the Workshop on BioNLP, Boulder, CO, USA, June 2009, pp. 107-110.

[16] S. Burr, "Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets", Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, 2004.

[17] S. Kulick, A. Bies, M. Liberman, "Integrated Annotation for Biomedical Information Extraction", Proceedings of HLT/NAACL 2014 Biolink Workshop, 2014, pp. 61-68.

[18] D. Mcclosky, M. Surdeanu, C. Manning, "Event extraction as dependency parsing for BioNLP", Proceedings of the BioNLP Shared Task Workshop, Portland, OR, USA, June 2011, pp. 41-45.

[19] M. Miwa, S. Pyysalo, T. Ohta, S. Ananiadou, "Wide coverage biomedical event extraction using multiple partially overlapping corpora", BMC Bioinformatics, Vol. 14, No. 1, 2013, pp. 175-175.

[20] C. C. Huang, Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure, and the future", Briefings in Bioinformatics, Vol. 17, No. 1, 2016, pp. 132-144.

[21] S. Riedela, D. Mccloskyb, M. Surdeanub, A. Mccalluma, C. Manning, "Model combination for event extraction in BioNLP", Proceedings of BioNLP Shared Task 2011 Workshop, Portland, OR, USA, June 2011, pp. 51-55.

[22] J.-D. Kim, T. Ohta, Y. Tsuruoka, "Introduction to the Bio- Entity Recognition Task at JNLPBA", Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Geneva, Switzerland, August 2014, pp. 70-75.

[23] D. Mcclosky, "Any domain parsing: automatic domain adaptation for natural language parsing", Providence, RI, USA. AAI3430199, 2010.

[24] E. Buyko, E. Beisswanger, U. Hahn, "The genereg corpus for gene expression regulation events- an

overview of the corpus and its in-domain and out-of-domain interoperability", Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 2010, pp. 1921-1921.

[25] J. D. Kim, T. Ohta, K. Oda, J. I. Tsujii, "From text to pathway: corpus annotation for knowledge acquisition from biomedical literature", Proceedings of the Asia-Pacific Bioinformatics Conference, Imperial College Press, 2008, pp. 165-176.

[26] J. Bjorne, T. Salakoski, "Generalizing biomedical event extraction", Proceedings of the BioNLP Shared Task 2011 Workshop, Portland, OR, USA, June 2011, pp. 183-191.

[27] S. Pyysalo, F. Ginter, J. Heimonen, "BioInfer: a corpus for information extraction in the biomedical domain", BMC Bioinformatics, Vol. 8, 2007, pp. 50-50.

[28] The LLL corpus, 2015.

[29] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, U. Leser, "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature", PLoS Computational Biology, 2010.

[30] A. Berger, S. Della Pietra, V. Della Pietra, "A maximum entropy approach to natural language processing", Computational Linguistics, Vol. 22, No. 1, 2016, pp. 39-71.

[31] Y. Tsuruoka, T. Yuka, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, J. Tsuji, "Developing a Robust Part-of- Speech Tagger for Biomedical Text", Proceedings of the 10th Panhellenic Conference on Informatics, 2015, pp. 382-392.

[32] T. Zhang, F. Damerau, D. Johnson, "Text chunking based on a generalization of Winnow", Journal of Machine Learning Research, 2012, pp. 615-637.

[33] J.-D. Kim, Y. Wang, N. Colic, S. H. Beak, Y. H. Kim, M. Song, "Refactoring the Genia event extraction shared task toward a general framework for IE-driven KB development", Proceedings of the Fourth BioNLP Shared Task Workshop, Berlin, Germany, August 2016, pp. 23-31.

[34] R. Mcdonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, P. White, "Simple algorithms for complex relation extraction with applications to biomedical", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, USA, June 2005.

[35] G. Li, K. E. Ross, C. N. Arighi, Y. Peng, C. H. Wu, K. V. Shanker, "MirTex: A text mining system for MiRNA-Gene relation extraction", PLoS Computational Biology, Vol. 11, No. 9, 2015, pp. 104-108.

[36] K. Yoshikawa, S. Riedel, T. Hirao, Asahara, Y. Matsumoto, "Coreference based event-argument relation extraction on biomedical text", Proceedings of the Fourth International Symposium on Semantic Mining in Biomedicine, 2010.

[37] D. Campos, S. Matos, J. L. Oliveira, "Biomedical named entity recognition: a survey of machine-learning tools", Theory and Applications for Advanced Text Mining, 2012, pp. 175-195.

[38] J. Berant, V. Srikumar, P. C. Chen, "Modeling biological processes for reading comprehension", Proceedings of the Empirical Methods in Natural Language Processing, Doha, Qatar, October 2014.

[39] L. Hirschman, M. Krallinger, A. Valencia, "Chemdner: The drugs and chemical names extraction challenge", Journal of Cheminformatics, Vol. 7, 2015, S1.

[40] J. D. Nguyen, M. Kim, T. Miwa, J. Matsuzaki, J. Tsujii, "Improving protein coreference resolution by simple semantic classification", BMC Bioinformatics, Vol. 13, 2012, pp. 304-304.

[41] Q. Leminh, S.N. Truong, Q.H. Bao, "A pattern approach for biomedical event annotation", Proceedings of the BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, Portland, OR, USA, June 2011, pp. 149-150.

[42] J. Gama, P. Brazdil, "Cascade generalization", Machine learning, Vol. 41, No. 3, 2014, pp. 315-343.

[43] A. Ozgur, D. Radev, "Supervised classification for extracting biomedical events", Proceedings of the Workshop on BioNLP, Boulder, CO, USA, June 2009, pp. 111-114.

[44] J. Lafferty, A. Mccallum, F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data", Proceedings of the Eighteenth International Conference on Machine Learning, June 2001, pp. 282-289.

[45] A. Airola, S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter, T. Salakoski, "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning", BMC Bioinformatics, Vol. 9, 2008, S2.

[46] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, U. Leser, "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature", PLoS Computational Biology, 2010.

[47] A. Majumder, A. Ekbal, S.K. Naskar, "Feature selection and class-weight tuning using genetic algorithm for bio-molecular event extraction", Proceedings of the NLDB, 2017.

[48] A. Majumder, A. Ekbal, S.K. Naskar, "Biomolecular event extraction using a stacked generalization-based classifier", Proceedings of the 13th International Conference on Natural Language Processing, 2016, pp. 55-64.

[49] K. Ting, I. Witten, "Issues in stacked generalization", Journal of Artificial Intelligence Research, Vol. 10, 2013, pp. 271-289.

[50] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, "A fast and elitist multi objective genetic algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation, Vol. 6, No. 2, 2002, pp. 182-197.

[51] H. Kilicoglu, S. Bergler, "Effective bio-event extraction using trigger words and syntactic dependencies", Computational Intelligence, Vol. 27, No. 4, 2011, pp. 583-609.

[52] J. Bjorne, F. Heimonen, Ginter, "Extracting complex biological events with rich graph-based feature sets", Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Boulder, CO, USA, June 2009, pp. 10-18.

[53] M. Miwa, R. Saetre, J.D. Kim, J. Tsujii, "Event extraction with complex event classification using rich features", Journal of Bioinformatics and Computational Biology, Vol. 8, No. 1, 2010, pp. 131-146.

[54] H. Kilicoglu, S. Bergler, "Syntactic dependency-based heuristics for biological event extraction", Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Association for Computational Linguistics, Boulder, CO, USA, June 2009, pp. 119-127.

[55] A. D. Casillas, K. D. Ilarraza, M. Gojenola, G. Oronoz, Rigau, "Using kybots for extracting events in biomedical texts", Proceedings of the BioNLP Shared Task 2011 Workshop, Portland, OR, USA, June 2011, pp. 138-142.

[56] S. V. Landeghem, B. De, Y. Baets, Y. D. Peer, Saeys, "High-precision bio-molecular event extraction from text using parallel binary classifiers", Computational Intelligence, Vol. 27, No. 4, 2011, pp. 645-664.

[57] D. Martinez, T. Baldwin, "Word sense disambiguation for event trigger word detection in biomedicine", BMC Bioinformatics, Vol. 12, 2011.

[58] D. Zhou, Y. He, "Biomedical events extraction using the hidden vector state model," Artificial Intelligence in Medicine, Vol. 53, No. 3, 2011, pp. 205-213.

[59] M. Miwa, P. Thompson, S. Ananiadou, "Boosting automatic event extraction from the literature using domain adaptation and coreference resolution", Bioinformatics, Vol. 28, No. 13, 2012, pp. 1759- 1765.

[60] F. Bjorne, T. Ginter, Salakoski, "University of Turku in the BioNLP'11 Shared Task", BMC Bioinformatics, Vol. 13, 2012, pp. 13.

[61] L. Qian, G. Zhou, "Tree kernel-based protein-protein interaction extraction from biomedical literature", Journal of Biomedical Informatics, Vol. 45, No. 3, 2012, pp. 535-543.

[62] J. Wang, Q. Xu, H. Lin, Z. Yang, Y. Li, "Semi-supervised method for biomedical event extraction", Proteome Science, Vol. 11, 2013, p. 17.

[63] K. Hakala, S.V. Landeghem, T. Salakoski, "EVEX: application of a large-scale text mining resource to event extraction and network construction", Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, August 2013, pp. 26-34

[64] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, "Biomolecular event trigger detection using neighborhood hash features", Journal of Theoretical Biology, Vol. 318, 2013, pp. 22-28.

[65] X. Liu, A. Bordes, Y. Grandvalet, "Biomedical event extraction by multi-class classification of pairs of

text entities", Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, August 2013, pp. 45-49.

[66] D. Campos, Q. C. Bui, S. Matos, J. L. Oliveira, "TrigNER: automatically optimized biomedical event trigger recognition on scientific documents", Source Code for Biology and Medicine, Vol. 9, No. 1, 2014.

[67] L. Li, Y. Wang, D. Huang, "Improving feature-based biomedical event extraction system by integrating argument information", Proceedings of the BioNLP Shared Task 2013 Workshop, Sofia, Bulgaria, August 2013, pp. 109-115.

[68] M. Bali, P. V. R Murthy, "Bio-Molecular Event Extraction Using Classifier Ensemble-of-Ensemble Technique", Data Management, Analytics and Innovation, Springer, 2021, pp.445-462.

[69] J. Xia, A. C. Fang, X. Zhang, "A novel feature selection strategy for enhanced biomedical event extraction using the Turku system", BioMed Research International, 2014.

[70] L. Hunter, Z. Lu, J. Firby, W. A. Baumgartner, H. L. Johnson, P. Ogren, V. K Cohen, "Opendmap: an opensource, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression", BMC Bioinformatics, Vol. 9, 2008, pp. 78-78.

[71] K. Yoshikawa, S. Riedel, T. Hirao, "Coreference based event-argument relation extraction on biomedical text", Journal of Biomedical Semantics, Vol. 2, 2011, S6.

[72] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, A. Valencia, "Overview of the protein- protein interaction annotation extraction task of Biocreative II", Genome Biology, Vol. 9, 2008, p. 4.

[73] P. Kordjamshidi, D. Roth, M. F. Moens, "Structured learning for spatial information extraction from biomedical text: bacteria biotopes", BMC Bioinformatics, Vol. 16, 2015, pp. 129-129.

[74] E. M. Voorhees, L. P. Buckland, "Proceedings of the Sixteenth Text Retrieval Conference", NIST, 2007, pp. 500-274.

[75] S. Riedel, D. Mcclosky, M. Surdeanu, A. Mccallum, C. D. Manning, "Model combination for event extraction in BioNLP", Proceedings of the BioNLP Shared Task 2011 Workshop, Association for Computational Linguistics, Portland, OR, USA, June 2011, pp. 51-55.

[76] J. Bjorne, T. Salakoski, "TEES 2.1: automated annotation scheme learning in the BioNLP 2013 shared task", Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, August 2013, pp. 16-25.

[77] Q. C. Bui, D. Campos, E. V. Mulligen, J. Kors, "A fast rule-based approach for biomedical event extraction", Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, August 2013, pp. 104-108.

[78] Y. Chen, L. Xu, K. Liu, D. Zeng, J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 1, 2015, pp. 167-176.

[79] T. Wang, Q. Kemao, "GPU Acceleration for Optical measurement", SPIE Press, 2017.

[80] J. Wang, H. Li, A.Y. Lin, H. Z. Yang, "Biomedical event trigger detection based on convolutional neural network", International Journal of Data Mining and Bioinformatics, Vol. 15, No. 3, 2016, pp. 195-213.

[81] Z. Han, J. Jiang, L. Qiao, Y. Dou, J. Xu, Z. Kan, "Accelerating Event Detection with DGCNN and FPGAs", Electronics, Vol. 9, No. 10, 2020, p. 1666.