Scalable Bio-Molecular Event Extraction System towards Knowledge Acquisition

Pattabhi RK Rao
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
pattabhi@au-kbc.org

Sindhuja Gopalan
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sindhujagopalan@au-kbc.org

Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

This paper presents a robust system for the automatic extraction of bio-molecular events from scientific texts. Event extraction provides information in the understanding of physiological and pathogenesis mechanisms. Event extraction from biomedical literature has a broad range of applications, such as knowledge base creation, knowledge discovery. Automatic event extraction is a challenging task due to ambiguity and diversity of natural language and linguistic phenomena, such as negations, anaphora and co-referencing leading to incorrect interpretation. In this work a machine learning based approach has been used for the event extraction. The methodology framework proposed in this work is derived from the perspective of natural language processing. The system includes a robust anaphora and coreference resolution module, developed as part of this work. An overall F-score of 54.25% is obtained, which is an improvement of 4% in comparison with the state of the art systems.

1 Introduction

Tremendous growth in the field of biomedical science has resulted in large amount of clinical and biomedical medical data. Primarily, the biomedical research largely focused on genome data analysis. Over the years, the application of new technologies to health care has resulted in voluminous data that includes structured and unstructured clinical notes, patient data, imaging data, etc. This growth also resulted in accumulation of large number of biomedical texts, i.e. medical literatures. It is important to extract useful information from these data to benefit the researchers for further findings. This requires the application of data driven approaches. Data mining involves the analysis and extraction of interesting patterns from large amount of data. In recent times the researchers are spending much effort on data mining for bioinformatics. The previous applications of data mining and machine learning (ML) to bioinformatics were on genetic data sets and phenotype data. Now it has been extended to text documents like clinical and biomedical data.

In the early days, the goal of natural language processing in biomedical domain was to populate the databases with biological information. This can be done manually, but requires lots of effort and is time consuming. Hence recognizing the named entities (NEs) using computational techniques could help in automatically populating the database with biological information. The extraction of the information like event or relation between biomedical entities will help the research community to compare the applicability of their works with others. Finding related literatures studying same biomedical entities is a crucial and challenging task. For example, there are lots of research publications related to "BRCA" gene. Unifying all studies about this gene helps the researchers to work on cancer therapy. The first step for accomplishing this task is extracting the biomedical named entities from literature and finding the events and relations between them.

Therefore, mining the literature and extracting the event between biomedical entities have lots of applications in bioinformatics. Event extraction from scientific texts in biomedical domain such as PubMed abstracts has attracted a lot of interest in the last decade, especially for those events involving proteins and other biomolecules. In the biomedical domain, an event refers to the change of state of one or more biomedical entities, such as proteins, cells, and chemicals. In the task of event extraction we

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 383–391

need to identify the types of the events and their arguments. Event arguments include event participants, which may be entities (e.g., proteins) or other events. This structured definition of events is associated with an ontology that defines the types of events and entities, semantic roles, and also any other attributes that may be assigned to an event. Examples of ontologies for describing bio-molecular events include the Genia Event Ontology. Consider the below Figure 1

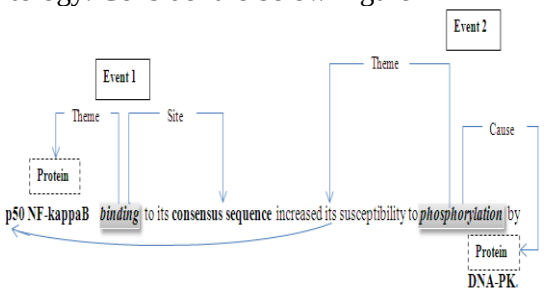

Figure 1: An Example for Event Occurrence in Biomedical text

The above Figure 1 shows two events, binding and phosphorylation. The first event belongs to event type binding, where the event has two arguments protein "p50 NF-kappaB", which is the theme and the second argument is the site "consensus sequence". The second event belongs to the event type phosphorylation and the arguments of this event is pronoun "its" which refers to the protein "p50 NF-kappaB", the theme of first event and protein DNA-PK, cause of second event. This example demonstrates the need for anaphora resolution in event identification.

By identifying the events we can extract information like gene-protein interactions, gene-chemical interactions and gene functions, etc. The BioNLP 2013 shared task on Genia Event extraction, has brought in more research groups to work in this area and has increased the research activity. Most of the systems which have participated in the BioNLP-ST 2013 Genia event (GE) extraction (Nédellec et al., 2013) have used the support vector machine (SVM) based pipeline. Two of the systems had used rule based approach. And another two had used hybrid approach where both rules and SVM have been used (Nédellec et al., 2013). In terms of use of pre-processing tools all the systems have used one of the deep parser tools such as McClosky-Charniak-Johnson Parser, Stanford Parser for syntactic processing. And some of the participating systems have also used external independent resource such as UniProt (Bairoch et al., 2005), IntAct (Kerrien et al., 2012), and CRAFT (Verspoor et al., 2012). Below we explain top two successful systems which have participated in the BioNLP-ST GE task. Both these systems have obtained an overall F-score of 50.97% and 50.74%.

Hakala et al., (2013) uses EVEX tool to extract the events. EVEX is a text mining resource built on top of events extracted from all PubMed abstracts and PubMed Central Open-Access full-text documents (Landeghem et al., 2011). Evex is built on top of "Banner" NER tool and "TEES" extraction tool. It uses SVM to re-rank the output of the EVEX resource output, sets a threshold score, below which the events are removed. The threshold score is obtained using a linear SVM regressor on each sentence. The results for event types "binding" and "regulation" are found to be lower and especially in "Methods" and "Captions" section of the documents.

Our work contributes to the application of data mining approach to biomedical data. This paper describes an event extraction system developed using ML approach and rich feature set including linguistic and biological domain motivated features. It has been observed from the participating systems in the BioNLP-ST 2013 that most of the systems have not used coreference resolution. Though the data had anaphora and coreference annotation, the systems had not exploited the annotations. In the present work we make use of the coreference annotations provided in the data. The use of coreference resolution has mainly improved the extraction of event type "binding".

The main contributions of this work are as follows:

1. We have developed a robust, scalable event identification system, which can be used for any of the biomedical domain documents. The developed system architecture is robust and portable for any biomedical text. The results obtained on the test data of the BioNLP ST 2013 GE task shows significant results comparable to the state-of the art.
2. We have developed a biomedical domain anaphora coreference resolution module for resolving the protein coreference relations. A general domain, robust anaphora coreference resolution module has been used and adapted (or customized) with the use of biomedical coreference annotations.
3. We have used open source tools such as "Genia tagger", CRF++ (Taku, 2005) for the development of the syntactic and semantic pre-processing. Thus this work is

easily implementable by other researchers.

In the following section we describe the corpora, features and the method used to develop the system. The results are discussed in Section 3. The paper ends with the conclusion and future works.

2 Method

The various approaches to event extraction task are rule based, dictionary based, ML and Hybrid approaches. This paper proposes a robust, scalable BioEventTag system developed by using graph-based ML technique Conditional Random Fields (CRFs) (Lafferty et al., 2001). This system is developed for the extraction of biological events. We have used CRF++ tool, an open source implementation of the CRFs algorithm. In this section, we present the experiments performed to extract the events from biomedical texts. First, the input text is syntactically preprocessed using Genia tagger. The pre-processing includes sentence splitting, tokenization, PoS tagging and chunking. Then in the next step semantic pre-processing is performed where the biomedical named entities (BNEs) are identified and anaphors in the document are resolved. For identifying the NEs we have used the biomedical named entity recognition (BioNER) system developed by (Gopalan et al., 2016). We developed an anaphora resolution system to resolve the anaphors. Finally we developed an event extraction system which contains two modules. First module identifies the event trigger and the second module extracts the event from the text. The system architecture is shown in Figure 2.

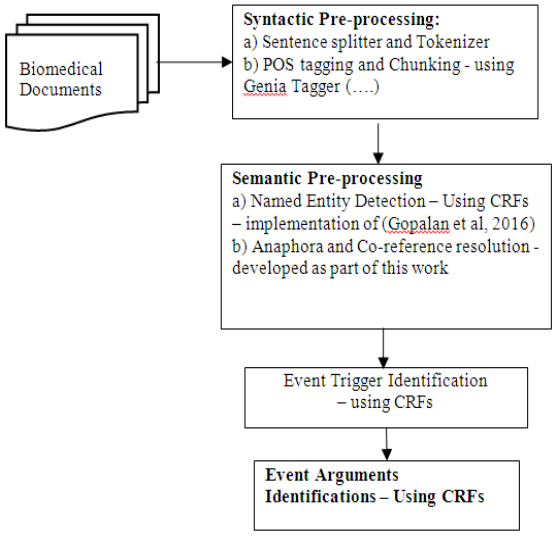

Figure 2: System Architecture

2.1 Corpora Collection and Analysis

We developed our system using a widely accepted dataset BioNLP-ST 2013 (Genia Event Task data). BioNLP-ST 2013 GE data was developed to evaluate the applicability of event extraction systems. The collection consists of 1210 titles and abstracts and 34 full papers from the Open Access subset of PubMed Central (Nédellec et al., 2013). Table 1 shows the corpus statistics. It is evident from the corpus statistics displayed in Table 1 that the majority of the events (65.86%) are "Regulation and Binding" event types. Thus handling of these two event types properly is very important to improve the system efficiency and performance. In these two types of events anaphora coreference resolution plays a significant role.

S.No	Description	Statistics
1	Number of Abstracts + full Papers	1210 + 34
2	Number of Words	2,63,133
3	Number of Proteins	16,427
4	Total Number of Events	9,364
5	Number of Anaphora and Coreference relations to Proteins	535
6	Number of Regulation and Binding Event types	6,168

Table 1: BioNLP Genia Event (GE) Shared Task Corpus Statistics

2.2 Feature Extraction

CRF++ is a general purpose tool and hence the feature template needs to be specified in advance. This file describes the feature used for training and testing. When the feature template is given, CRF++ automatically generates a set of feature function. The challenge in developing an event extraction system using ML techniques lies in designating the striking features and designing of feature template. We have used window size of 5 for this work. We describe in detail the features used in developing our system.

Lexical features and **Syntactic features** such as word, Parts of Speech (PoS) and chunk are used. PoS help in disambiguating the sense of the word in a sentence. PoS is an important feature for extracting the events as most of the arguments of an event are proper noun and event trigger belongs to noun and verb category. Hence PoS is a key feature for event extraction task. Most of the event trigger and arguments are descriptive i.e., they occur as a phrase. Hence

chunk tag will help in argument and event trigger extraction.

Morphological Patterns: Prefix/suffix is used as one of the features in our work. For example, an event trigger like "phosphorylation", has suffix '-ation', which means action or process. Prefix/suffix of a token helps to boost the performance of the system.

Biomedical Named Entities: BNEs are used as features in our work. BioNER is the task of extraction of BNEs like gene, protein, chemical etc. from biomedical text. From Figure 1, we can observe that the arguments of the events are BNEs and hence BNEs are useful features for argument identification task.

The combination of these features is used to develop the template feature. The template file sets up which features to use while running CRFs. Each line in the template file represents one template. The template is represented as %x[row,col], where "row" specifies the relative position from the current token and "col" represents the absolute position of the column.

2.3 Experiments

In this work we have followed two step approach, first the event trigger is identified and then the event arguments. One important semantic pre-processing module has been introduced in this work. We have developed Anaphora and Coreference resolution module as part of semantic pre-processing. This is important to resolve the anaphoric entities such as "these proteins", "it" which refer to proteins, chemicals etc. After, incorporating the features, the system was trained with the training corpus. We extracted the distinctive features to build the language models based on conditioned features. Finally, by using these language models, NEs in the testing corpus are automatically labeled. The experiments performed are detailed in this section.

Syntactic Pre-Processing: The syntactic preprocessing of the data is performed using Genia Tagger (Tsuruoka et al., 2005), where the data is split into sentences and tokenized and then PoS and chunk tags are added. The performance of this tool for PoS tagging is 98.26% accuracy and for chunking, the F-score obtained is 88.9% for Noun Phrases and 95.2% for Verb phrases (Kang et al., 2011).

Semantic pre-processing: The semantic preprocessing of the data includes named entity tagging and anaphora resolution. As the event in biomedical text is established between the BNEs, the identification of BNEs is important. Similarly as described in Section 1, resolution of anaphora is an essential step for event extraction that helps in improving the system's performance.

Biomedical Named Entity Recognition: The BNEs are identified using the system developed by (Gopalan et al., 2016). This portable system is developed on three data sets, BioNLP/NLPBA 2004 dataset; BioNLP-ST 2013 (pathway curation task data) and BioCreative 2013 CTD track data using ML approach. A rich feature set including linguistic features and domain-specific features were used to develop the system. For BioNLP-ST corpus they obtained F-score of 83.73%. The BNEs belonging to calss simple chemical, gene or gene product, complex and cellular component are identified. We used this system to identify the named entities from our corpus. Named entities is one of the features used for event argument identification. After identifying the entities, the resolution of anaphora is performed.

Biomedical Anaphora and Co-reference resolution module: Anaphora is a compound word consisting of the words "Ana" and "phora". "Ana" refers to back, upstream or back in an upward direction. "phora" means the act of carrying and denoted the act of carrying back stream. Anaphora is a type of expression whose reference depends upon another referential element. Reference is made based on the preceding part of the utterance. It is the cohesion which points back to some previous items. "The pointing back" is called an anaphor and the entity to which it refers is antecedent. The process of determining the antecedent of anaphor is called as anaphora resolution. Anaphora resolution in discourse is the task or process of identifying the referents of expressions which we use to denote discourse entities, i.e., objects, individuals, properties and relations that have been introduced and talked about in the prior discourse. Biomedical texts differ significantly from other text genres such as newspapers and fiction writing. In biomedical texts, much background knowledge is required for the reader to understand the relation between the entities mentioned in the text. This is a common aspect of scientific papers.

One of the common problems in biomedical texts is a gene and the protein it encodes share the same name, causing some ambiguity in the text when the context does not provide enough information to determine whether the writer is talking about the gene or the protein. Though there are writing conventions to avoid this ambiguity, it is common, however, that authors do not

follow these conventions properly. Other common issue is protein or gene names may coincide with common English words, e.g. for (symbol for foraging). These sources of ambiguity create challenges to a system for automatic detection of entities and events.

The distribution of different types of noun phrases in biomedical articles differs from the distribution in other general text. Pronouns are very rare, accounting for about 3% of noun phrases; whereas proper names, acronyms are very frequent, giving mentions of genes, proteins and names of other BNEs. In the WSJ Newswire corpus the pronouns are 4.5% and in fiction part of the brown corpus the pronouns are 22%. Another aspect in the pronouns distribution in the biomedical texts is it has more plural pronouns such as "these proteins", "them", "those proteins". This makes the task of anaphora and coreference resolution more challenging. It is observed that in biomedical texts entities are commonly referred to using non-pronominal noun phrases, like proper nouns, acronyms or definite descriptions. Hence there is a need to focus on these noun phrases (NPs) for a good event extraction engine. The occurrence of acronyms and NPs which have part-of relationships, linking those in the co-reference chain is a challenging aspect in biomedical coreference resolution.

Though there are many Anaphora and coreference resolutions systems developed for general domain, there are very few works on anaphora and coreference resolution in the biomedical domain. Castano et al., (2002) developed a salience-based system for anaphora resolution which uses UMLS Semantic Network to obtain semantic information. Gaizauskas et al., (2003) developed PASTA system, which implements an inference-based coreference resolution module. Yang et al., (2004) developed a supervised ML approach for anaphora resolution and evaluated it on a portion of the GENIA corpus. Gasperin et al., (2009) developed a statistical anaphora resolution system for biomedical domain. They have tested their system on various corpora.

Here in this work, a general Newswire text anaphora engine is customized to adapt to the biomedical domain. So here anaphora and coreference module is developed using a hybrid approach. An implementation of a general anaphora and coreference resolution engine as described in Lalitha Devi et al., (2011) is done here. The main difference in the implementation is the corpus used for training. Lalitha Devi et al., (2011) use Newswire text documents, whereas in this implementation, the anaphora and coreference annotations provided in BioNLP-ST 2013 GE task have been used. The results obtained from this engine are post processed with rules specific to biological domain. In the post processing stage Genia ontology is used to provide the required world knowledge to resolve the linking of acronyms, for improving the resolution of acronyms and definite descriptions. In Figure 3 the architecture of Anaphora and coreference module is shown in detail.

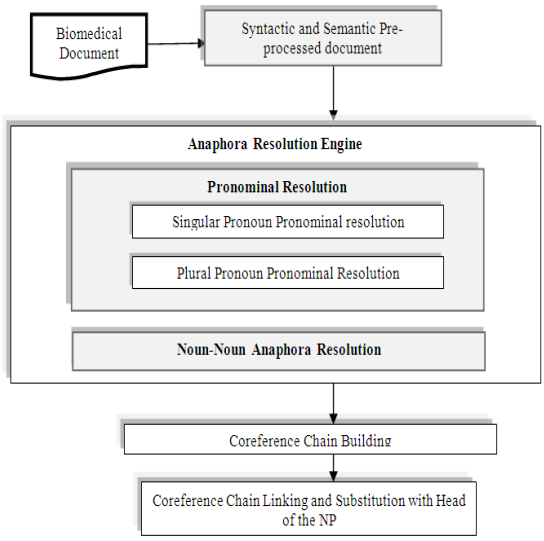

Figure 3: System architecture for Anaphora Resolution

The features used are same as described in the Lalitha Devi et al., (2011). Along with those features, two more features specific to biological domain are added. The new features are

1. Biological Entity type matching: 'yes' if anaphor's and candidate's biological entity type match, 'no' otherwise.
2. Is Entity type a gene or Protein? 'Yes' if the anaphor entity type or candidate entity type is gene or protein, 'no' otherwise. This feature is mainly to distinguish which pairs can hold BNE relations, because most of the event types have arguments as proteins or genes.

This module has been evaluated separately to ascertain its efficiency as a standalone engine. This has been tested on the gold anaphora annotations provided in the test partition of the GE task of the BioNLP 2013 shared task. We have obtained a precision of 55.35%, recall of 58.36% and F-score of 56.86%.

Event extraction: The whole task of event extraction is divided into two sub-tasks. First the event trigger is identified and then the event arguments are extracted. The event extraction sub task includes two phases, event start identifica-

tion and event end identification. We have used the training and development partitions of the GE task data for training and the test partition has been used for testing the system. The experiments performed for event trigger identification and event extraction are described below.

Event Trigger Identification: Event trigger is an important feature for extraction of event from a document. The features used for event trigger identification includes lexico-syntactic features like words, PoS, chunk and morphological patterns. The event trigger for event relation includes noun phrases containing action terms like "regulation", "interaction", "phosphorylation", "expression" etc. In some cases, the event trigger is verb phrases like "activates" that belongs to event type "positive regulation". Hence syntactic features like PoS and chunk acts as prime features for identification of event triggers. In addition to lexical and syntactic features, we have also used another biomedical domain specific feature, "trigger indicator". Trigger indicator feature includes biomedical domain specific verbs such as "binds", "inhibit" etc. and biomedical key terms like "translocation", "methylation" etc. This feature has a Boolean value "true" if domain specific verbs or key terms are identified in the current word, else "false". For event trigger identification the data is first preprocessed and features are extracted. After extracting the features, the language model is built. The identification of event trigger is followed by the identification of event arguments.

Event Argument Identification: The event extraction task includes extraction of event and its arguments. For extraction of event arguments, the event start and event end is identified. The event trigger is identified in the first task and the sentences with event trigger are given as input to the event argument extraction module. The features for event boundary identification are word, PoS, chunk, event trigger and BNEs. The arguments for the events are BNEs. Hence, giving weightage to BNEs that occur before or after the event trigger helps in the identification of argument boundaries. Using these features the language models are built for event start boundary and event end boundary. These models are used for identifying and extracting the event of a text.

The event may have one argument or multiple arguments. In case of events like "gene expression", there will be one argument "theme". Whereas in case of events like "binding" and "regulation", there will be more than one arguments such as "theme", "cause" and "site". Consider the Example 1 given below.

Example 1:

Methyl-CpG-binding proteins (MBPs) are thought to **inhibit** the **binding** of ***transcriptional factors*** to the ***promoter***.

In this Example 1 there are two events "negative regulation" and "binding". The event triggers are "inhibit" for negative regulation and "binding" for binding event. The arguments for "negative regulation" event is the theme "binding of transcriptional factors", which again is an event, the cause "Methyl-CpG-binding proteins" and the site "promoter". This event has three arguments. The second event is "binding" and the arguments of this event are the theme "transcriptional factors" and the site "promoter". For this event there are two arguments. The simpler events mostly have one argument.

From this example we also observe that the arguments of first event and second events are overlapping and importantly the second event as a whole is one of the argument of first event. Both the events share same arguments and hence the argument boundaries overlap. In these cases we have processed the events separately i.e. when there is more than one event in a sentence; each event is processed one by one, while developing the models.

The event arguments are actually relations between the entities. Thus the event argument identification is modeled as the identification of argument spans for each argument of the event trigger. The basic assumption is that each event will either have an explicit or an implicit event trigger. Event argument span identification is split into four sub-phases for identification of each boundary of each argument, i.e., the identification of Arg1's two boundaries and Arg2's two boundaries. Four language models were built for this purpose and Arg2-START, Arg1-END, Arg1-START and Arg2-END were identified in series, in that order. The output at each sub-phase was fed as input to the next sub-phase. In other words, in each sub-phase, the previously identified boundary is also used as a feature along with the features explained in Section 2.2. The choice of the order of identification of bounds was made with the idea that it is easier to first find the boundaries that are in close proximity to the cause-effect marker – Arg1-END and Arg2-START. Between these two, Arg2-START was chosen first, arbitrarily. The same holds for the choice of Arg1-START to be the third boundary. The arguments need not be always adjacent to

the marker. Sometimes, the arguments can be in the same sentence as the event trigger as shown in Example 1. Sometimes, one of the arguments is in the sentence immediately preceding that of the event trigger.

3 Results and Discussion

This section describes the performance of our system in terms of Precision, Recall and F score. Precision is the number of NEs correctly perceived by the system from the total number of NEs identified, Recall is the number of NEs correctly detected by the system by the total number of NEs contained in the input text and F-score is merely the harmonic mean of precision and recall.

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F score = (2 × Recall × Precision / (Recall + Precision))

Where, TP means true positives, FN means false negatives and FP means false positives.

We evaluated our system on test partition of GE task data. The overall result and results achieved by the system for each event type are demonstrated in Table 2 and Table 3. First, we developed the system for event extraction, without resolving the anaphors. Then we improved the performance of the system by resolving the anaphors. We obtained F-score of 75.12% for simple events, 62.11% for Binding Events & Protein Modification Events, 35.31% for regulation events and 49.27% for all event types without resolving the anaphors. After resolving the anaphors we observed that there is a significant increase in the performance of the system for identification of binding and modification events and regulation events.

Event Types	Precision	Recall	F-score
Simple Events	78.65	71.89	75.12
Binding & Protein Modification Events	66.36	58.54	62.11
Regulation Events	41.43	30.77	35.31
Overall	60.15	41.73	49.27

Table 2: Results for event extraction- Without Anaphora & Coreference resolution

Event type	Precision	Recall	F-score
Simple Events	78.75	71.94	75.76
Binding & Protein Modification Events	69.87	61.67	65.51
Regulation Events	46.15	35.42	40.08
Overall	67.15	47.73	54.25

Table 3: Results for event extraction-After Anaphora & Coreference resolution

Although simpler events achieve good results in event extraction task, the extraction of events such as binding and modification events and regulation events is still difficult. We have made an approach to improve the results of these complex events by resolving anaphora and co-reference. There are 445 anaphora relations in binding and regulation event types. With the help of our anaphora resolution engine we were able to identify the referents of 254 anaphor relations correctly. The anaphor relation consisted of pronominal anaphors such as them, its, they etc. and noun-noun anaphors such as "aforementioned cytokines", these proteins" etc. Consider the below Example 2

Example 2

After 5 days, supernatants were collected and the secretion of ***IFNgamma, IL4 and IL2*** were measured by ELISA. Samples from both negative controls had no detectable production of the ***aforementioned cytokines***.

In Example 2, the event trigger is "production" and the event is "gene expression". The argument for the event is "aforementioned cytokines", but this refers to IFNgamma, IL4 and IL2. This is an example for noun-noun anaphora relation. The main objective of this task is to identify the protein involved in the event. If we do not resolve the noun-noun anaphor "aforementioned cytokines", we will not be able to identify the protein names. Hence resolving the anaphors helped in the improvement of regulation and binding events. To know the significance of each features we conducted experiments to check the performance of individual features. The results for performance of individual features are shown in Table 4.

Table 4: Results for Individual features

Features	Precision	Recall	F-score
Lexical feature	37.87	23.13	28.53
Lexical feature +Syntactic features	49.46	28.79	36.80
Lexical feature +Syntactic features +Event trigger	60.56	37.01	45.94
Lexical feature +Syntactic features +Event trigger + BNEs	60.45	41.65	49.31
All above features + Anaphora	67.15	47.73	54.25

For event extraction we have used lexical feature, syntactic features such as PoS and chunk, event trigger and biomedical entities. When lexicon is used as feature we obtained precision of

37.87%, recall of 23.13% and 28.53% F-score. A window size of 5 is used. Then we used syntactic features along with the lexical feature. Since PoS and chunk plays a key role in extraction of event trigger and arguments, we observed that there were significant improvement in precision and recall of the system. There was a significant increase in F-score of about 8.27%. Then we used event trigger as feature along with lexical and syntactic feature. Event trigger is very important feature in event extraction task as it signals the presence of event in a text. We obtained good increase in performance with precision of 60.56%., recall of 37.01% and F-score of 45.94%. Then we used BNE features along with the other features. BNEs are main features for argument identification of an event. This feature helped in heightening the system's performance with increase in F-score of about 3.37%. We obtained an increase of 4.94% after resolving the anaphors. The evaluation results show that the system is comparable to state of art system.

4 Conclusion

This paper described an event extraction system designed using the ML approach CRFs, with rich feature set. We have evaluated our system on test partition of GE task data and showed that the system is comparable to state of art system. The performance of the system based on individual feature is outlined and have also exhibited that the system render good performance by resolving the anaphors.

Reference

John Lafferty, Andrew McCallum and Fernando C. N. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data, *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, USA, 282–289.

Kudo Taku. 2005. *CRF++, An Open Source Toolkit for CRF* [online] http://crfpp.sourceforge.net (accessed 3 January 2013).

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo and Pierre Zweigenbaum. 2013. Overview of BioNLP Shared Task 2013, *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 1–7.

Sindhuja Gopalan and Sobha L. Devi. 2016. BNE-Miner: mining biomedical literature for extraction of biological target, disease and chemical entities, *Int. J. Business Intelligence and Data Mining*, 11(2):190–204.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh. L.S.L. (2005) 'The universal protein resource (uniprot)', *Nucleic Acids Research*, 33(1):154–159.

Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C. Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard and Henning Hermjakob. 2012. The intact molecular interaction database in 2012, *Nucleic Acids Research*, 40(1):841–846.

Karin Verspoor, Kevin B. Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner, Michael Bada, Martha Palmer and Lawrence Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools, *BMC Bioinformatics*, 13(1): 207.

Jose Castano, Jason Zhang and James Pustejovsky. 2002. Anaphora resolution in biomedical literature, *Proceedings of International Symposium on Reference Resolution for NLP 2002*, Alicante, Spain.

Robert Gaizauskas, Demetriou, G., Artymiuk, P.J., and Willett, P. (2003) Protein structures and information extraction from biological texts: the PASTA system, *Bioinformatics*, 9(1):135-143.

Xiaofeng Yang, Jian Su, Guodong Zhou and Chew L. Tan. 2004. An NP-cluster based approach to coreference resolution, P*roceedings of COLING 2004*, Geneva, Switzerland, 226–232.

Jari Bjorne and Tapio Salakoski. 2013. TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task, *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 16–25.

Kai Hakala, Sofie V. Landeghem, Tapio Salakoski, Yves Van de Peer and Filip Ginter. 2013. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction, *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, 26-34.

Sobha L. Devi, Pattabhi R. K. Rao, Vijay S. Ram, Malarkodi C. S, and Akilandeswari A. 2011. Hybrid Approach for Coreference Resolution, *Proceedings of 15th Conference on Computational Natural Language Learning: Shared Task*, Portland, Oregon, 93-96.

Sofie V. Landeghem, Filip Ginter, Yves Van de Peer and Tapio Salakoski. 2011. Evex: A pubmed-scale resource for homology-based generalization of text mining predictions, *Proceedings of BioNLP 2011 Workshop*, Portland, Oregon, USA, 28–37.

Caroline V. Gasperin. 2009. Statistical anaphora resolution in biomedical texts. A Technical report based on a dissertation titled Statistical anaphora resolution in biomedical texts, University of Cambridge.

Ning Kang, Erik M. van Mulligen and Jan A. Kors. 2011. Comparing and combining chunkers of biomedical text, *Journal of Biomedical Informatics*, 44(2):354–360.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text, *Lec-*