2014

Combination of Rule-based and Machine Learning for Biomedical Event Extraction

Xuan-Quang Pham
HoChiMinh City University of Science,Vietnam, pxquang@fit.hcmus.edu.vn

Bao-Quoc HO
HoChiMinh City University of Science,Vietnam, hbquoc@fit.hcmus.edu.vn

11P. A Combination of Rule-based and Machine Learning for Biomedical Event Extraction

Xuan-Quang Pham
HoChiMinh City University of Science,Vietnam
pxquang@fit.hcmus.edu.vn

Bao-Quoc HO
HoChiMinh City University of Science,Vietnam
hbquoc@fit.hcmus.edu.vn

Abstract

This paper describes the method for biomedical event extraction. The biomedical events occurs in relative to biomedical concepts (objects) as proteins, genes. In this work, we try a hybrid method to identify given event types relative to a given set of proteins in biomedical text. The approach combines rule-based and machine learning. A Set of rules is built based on event triggers, and a set of features is selected to use for machine learning algorithm. Our system consists of four main phases: preprocessing, trigger detection, event detection and post-processing. These phases are developed based on UIMA[1] framework. This work is continuous of our work for BioNLP2013 Shared Task[2]. The final result obtains 36.60 f-score.

Keywords

Biomedical event extraction, Information extraction, Text mining, Machine Learning.

1. Introduction

Biomedicine is attracting a lot of attentions from the research community. It is risen from the demand of automatic processing of a large volume of literature, papers or documents. The fundamental task is biomedical entity recognition. These entities are proteins, genes or disease names (Saha et al. 2012). Another research branch focuses on binary relations extraction of entities such as protein-protein interactions (Asur et al. 2011), gene-disease and drug-protein relations. Recently, more complex relations are targets, for example, co-reference resolution and event extraction which are more informative and useful.

All above problems have been introduced in many conferences and shared tasks. Our work is prepared to deal with the challenge tasks of the BioNLP Shared Task (BioNLP-ST).

The BioNLP-ST is a series of efforts to promote a community wide collaboration towards fine-grained information extraction (IE) in biomedical domain. Through two previous events, its tasks are expanding and generalizing. There are six main tasks in BioNLP2013, which are hot topics and necessary to support biologists. They include GENIA Event Extraction, Cancer Genetics, Pathway Curation, Gene Regulation Ontology, Gene Regulation Network and Bacteria Biotopes (Nédellec et al. 2013). Due to the fact that the scope of these topics is large, we choose only one topic, the GENIA Event Extraction (GE) task (Kim et al. 2013).

[1] http://uima.apache.org/
[2] http://2013.bionlp-st.org/

In the GE task (2013), there are 6 main event classes and 7 sub classes. Those events which are related to transcription factors in human blood cells can be simple or complex. The simple event is just binary relations while the complex event consists of more than one binary relation. The complex events provide important information for modeling biological systems. When identifying the events, it is required to detect event type together with the event trigger (the sign to recognize an event /anchor) and primary arguments of each event. The event may have one to a few arguments and the more complex one may have another event as its argument.

For GE task, we apply a hybrid approach in which rule-based and machine learning are combined.

The remainder of this paper is organized as follows. The next section presents related work. Section 3 explains our proposed approach. Section 4 discusses the experimental results and evaluation. Finally, we present the conclusion in section 5.

2. Related Work

General model for biomedical event extraction deals with associations between arguments (proteins, entities or other events) and anchors (event trigger). They may be in a single sentence or cross sentences. Extracting relation or interaction of entity pairs is simpler than event extraction, but there are similar characteristics between them. Therefore, features and methods of relation extraction are developed and adjusted for event extraction.

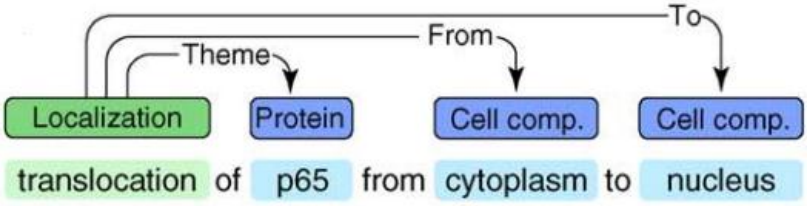


Figure 1: An Example of biomedical events

In event extraction, the rule-based approach is used in (Kaljurand et al. 2009; Kilicoglu & Bergler 2011), machine learning (ML)-based (Bjorne et al. 2009; Miwa et al. 2010) and hybrid method (Riedel & McCallum 2011; Riedel et al. 2011). Recently, (Riedel & McCallum 2011) present an approach based on optimization of scoring sets of binary variables.

According to the summaries for BioNLP-ST 2009 and 2011 (Kim et al. 2009; Kim et al. 2011), the results of ML-based method are better than the rule-based method. However ML is nontrivial to apply. The summary also indicates that high precision, for simple events, can be achieved by rule-based approach. In 2013, we participated on BioNLP-ST with a hybrid approach (Quang et al. 2013). This work is presented in the next section.

3. Proposed Approach

In this paper, a hybrid approach for extracting event from biomedical literature is proposed. Our approach combines the strengths of both semantic rule-based and machine learning classification. The main idea of the approach is usage of linguistic information as word morphology, syntactic graph as syntactic patterns and dependency graph of sentence as features to classify event triggers and events by itself. The proposed method consists of two phases: using a set of rules to extract events correctly recognized by the rules and then, using a set of features to identify more others events. First, it uses linguistic information from

syntactic graph to create two kinds of rules. The first kind of rules will check on the path on syntactic graph. The second kind is based on POS tagger label which is applied for event triggers in noun and verb form (Bui & Sloot 2011). Second, we reused the features as: shortest dependency path between predicted trigger and arguments (Razvan et al. 2005), some important features such as: word, n-gram, word frequency, dependency features (Amami et al. 2012) and proposed new features as token contains both protein and trigger, and detected events by rule-based for classification by machine learning method. Finally, the both rule-based and machine learning are combined. The details are presented in section 3.3.

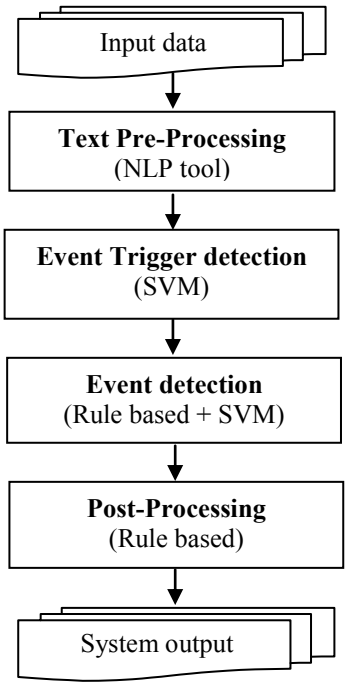


Figure 2: The main phases of the system

The overall architecture of the system is shown in Fig. 2. The main phases of our event extraction system are: pre-processing, event trigger detection, event extraction and post processing.

3.1 Text Pre-Processing

At first, the input documents are processed by NLP tools. These tasks include: sentence splitting, tokenization, POS tagger and dependency parsing. The linguistic information will be used to recognize event trigger and for each candidate event, its arguments will be identified.

In this phase, the list of given proteins is used to select the sentences which may contain events related to given protein.

At the end of this phase, a set of candidate sentences are selected to process further.

3.2 Event Trigger Detection

Because there are lots of ambiguity when recognizing a predicted event trigger, the system firstly have to simplify and restrict the number of tokens needed to classify, some heuristic restrictions are applied. It just considers those tokens having part-of-speech (POS) tags of

noun, verb and adjective. Then, we try to use a simple dictionary built from training data to check whether or not those tokens are triggers set and combine with machine learning approach. The problem is which features could be used to identify an event trigger. We analyzed the features are used in the literatures (Bjorne et al. 2009) and the data on training data set to select the features. The feature set in our system includes:

- Trigger words: word (base on a given list of trigger word), POS tagger label
- trigger word context (-2, +2): word, tagger label
- Special characters: contains "; /", number
- Known protein: contains a protein, ex: IL-10-expressing in which IL-10 is a protein and this word represent an event trigger
- Linguistic dependency with a protein in the same sentence: noun-noun dependency

The result archive on the development test is Precision: 69.89; Recall: 79.15, F-Score: 74.23

3.3 Event Detection

The next step relies on detected event triggers to identify events. There are two kinds of event. The basic event which concern the event trigger and the protein and the complex event which combine many events. At first, the rule-based are used to extract the events, and then the machine learning are used.

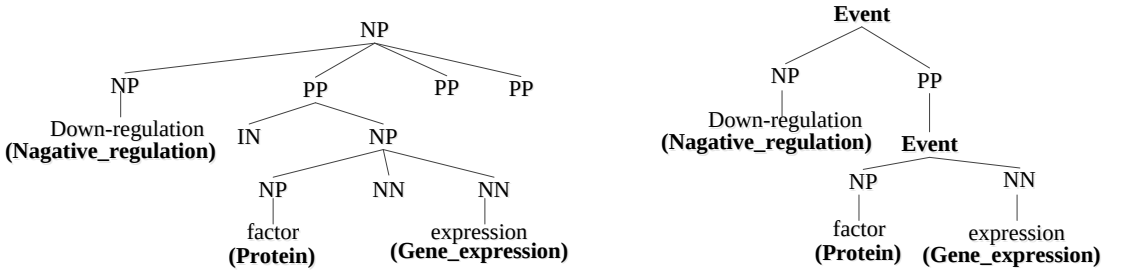
3.3.1 Rule-based task

In this important part, an event is formed by combining a trigger with appropriate arguments in the right context. According to the description of the shared task, nine targeted event have core arguments and additional arguments. However, the final evaluation uses the method which considers the core arguments as the primary scores. Therefore, at first, only the core arguments are focused, the additional arguments will be process in the future development. In addition, events which cross sentences are not considered.

The system divide nine events into three groups depended on characteristics of their core arguments. The first group is simple event (5 types of event) which has only one argument and the argument is protein. The second is Binding Event, consisting of unlimited the number of core arguments, but arguments must be protein too. The last includes remaining three types. Their core argument can be either Protein or another Event. Therefore, the third class may form a nest structure. The method will be lightly different for each group.

Arguments of an event are combined of Protein, Entity and Event Trigger in one sentence. Due to the vast amount of the combination, some restrictions are set. The combination has to contain at least one Protein (for three groups) or one Trigger (for the third group) and total is less than four. We also add a threshold of the difference between arguments' depth and the anchor's depth in the parse tree.

In this step, the two kinds of rules are applied. Both of them use linguistic information from syntactic graph. They are run separately; finally the two result sets are combined.



(a) The parse tree contains event trigger and protein

(b) The prune parse tree with assigned argument's concept

TOP (*Event* (**Nagative_regulation** *Event* (Protein **Gene_expression**)))

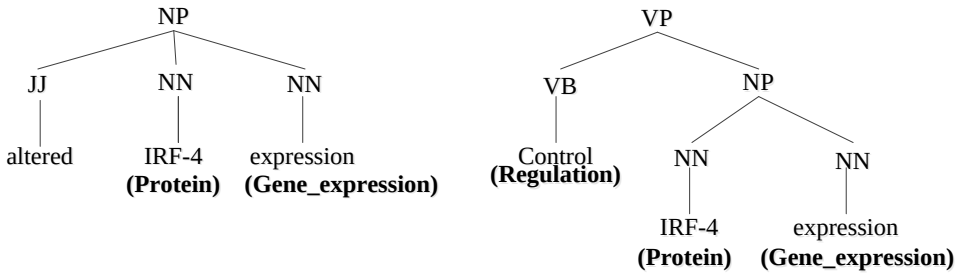
TOP (*Event* (Protein **Gene_expression**))

(c) Two string-form patterns from the tree above

Figure 3: Sample patterns of the first kind of rule.

The first kind of rule-based on syntactic graph will be learned from training corpus. The original parse tree of each sentence containing at least one trigger is retrieved. Nodes which only one branch are pruned and kept the top node to retain the most important parts. Concepts of candidate arguments (name role) and the trigger are assigned to appropriate tree-nodes according to their spans in the text. Next, we find the closest parent of all arguments. The patterns are the string form of the sub-tree of the modified parse tree (shown in Fig. 3). We will use this set of pattern to recognize the event from test corpus.

The second kinds, is based on POS tagger lable of nodes in syntactic graph. This idea is inspired from extracting protein-protein interactions (Bui & Sloot 2011), we construct some patterns connecting arguments and triggers. There are two kinds of patterns: noun phrases (NP) and verb phrases (VP). Each phrase has to have one trigger and at least one Protein. In the case of the NP, it contains two nouns without other phrase or it includes a preposition phrase (PP) and the trigger has to be the head of this NP. The second pattern, we find a VP which is a direct parent of the trigger. Two examples of satisfied patterns in NP and VP are shown in Fig. 4. If there is a Protein in those phrases, we annotate an Event with the trigger and the Protein as core argument.



(a) Noun phrase form

(b) Verb phrase form

Figure 4: Two kinds of form in the second kind of rule

3.3.2 Machine learning tasks

The SVM classifier with a linear kernel is used to classify the relation between two objects (trigger/protein or trigger/event) in the sentence to given event class. Two proposed features

include: the token contains both event trigger and protein, and detected events collection by the rule-based. In addition, the special feature is shortest dependency path between predicted trigger and arguments (proteins or events) (Razvan et al. 2005) and the other important features (Amami et al. 2012) are:
Element features: trigger/argument word, trigger/argument type and trigger/argument POS.
N-gram features: n-grams of dependencies, n-grams of words and n-gram of consecutive words representing governor-dependent relationship.
Frequency features: length of the shortest path between trigger and argument (protein or event), number of arguments and event triggers per type in the sentence and the frequency of trigger in corpus.
Dependency features: Directions of dependency edges relative to the shortest path, types of dependency edges relative to the shortest path.

3.4 Post processing
In this phase, the complex events will be recognized based on the detected events. In the previous phases, the complex event may be unrecognized because we have analyzed local part of syntactic tree. Now, the system try to combine the events on the global view (whole of the tree).
Then another task is to remove the duplicated events from two approaches (rule and machine learning).

4. Experimentation

4.1 Implementation
We evaluated the performance of our system by using the evaluation system for the development[3] and test[4] data set which provided by share task organizers. Errors were also analyzed on the development data set.

The dataset are run and provided for participants and the McClosky-Charniak-Johnson Parser[5] tool is chose for syntactic analyses. That parser is improved from Stanford parser with self-trained biomedical model. After passed throughout text pre-processing, continuously the trigger and event detection. The creating shortest path use Breadth-first-search (BFS) algorithm which is in JAVA. We implement JAVA method to extract features for event detection sub tasks, i.e., trigger detection, between two objects relation. For SVM classification, we use the LIBSVM[6] software

4.2 Performance
The table 1 shows the latest results of our system through the evaluation service of the organization. We achieved f-score rate is 36.60. The main reason of uneven scores between simple events is differential distribution of simple events. The lowest results in the group class of regulation because two reasons. First, they depend on the result of simple events. Second, the combination of arguments and union of two approaches cause the number of wrong answer increases significantly. One weak point of the system is the negative events detection. We need to improve in the future.

[3] http://bionlp-st.dbcls.jp/GE/2013/eval-development/
[4] http://bionlp-st.dbcls.jp/GE/2013/eval-test/
5 http://bllip.cs.brown.edu/resources.shtml
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Event Class	*Recall*	*Precision*	*F-score*
Gene_expression	76.41	66.34	71.02
Transcription	31.68	55.17	40.25
Protein_catabolism	57.14	61.54	59.26
Localization	25.25	52.08	34.01
Phosphorylation	74.38	63.30	68.39
Binding	37.24	31.39	34.07
Protein_modification	0.00	0.00	0.00
Ubiquitination	0.00	0.00	0.00
Acetylation	0.00	0.00	0.00
Deacetylation	0.00	0.00	0.00
Regulation	9.03	11.98	10.30
Positive_regulation	18.58	23.23	20.65
Negative_regulation	23.19	31.52	26.73
Event Total	**34.50**	**38.97**	**36.60**

Table 1: Evaluation Results on Test set

5. Conclusions

In this paper, the combination of rule-based and machine learning approach is used to extract biomedical events. The rule is based on linguistic information as patterns of syntactical tree or patterns of POS tagger lable. The selected features for machine learning are word form, context, frequency, dependency graph. The approach is evaluated on data set of BioNLP Shared Task 2013. It archives F-score 36.60, Precision 38.97, Recall 34.50. The result is not so high but it encourages the efforts to use of hybrid approach.

In the future, it need to be enrich the rule set by experts and figure out more effective features for machine learning to improve the performance.

References

Amami, M., R. Faiz and A. Elkhlifi (2012), "A framework for biological event extraction from text," Copyright 2012 ACM 978-1-4503-0915-8/12/06. WIMS' 12 June 13-15, Craiova, Romania.

Asur, S., D. Ucar, S. Parthasarathy (2007), "An ensemble framework for clustering protein-protein interaction networks," Bioinformatics, 23(13): i29-i40.

Bjorne, J., J. Heimonen, F. Ginter, A. Airola, T. Pahikkala and T. Salakoski (2009), "Extracting Complex Biological Events with Rich Graph-Based Feature Sets," In Proceedings of BioNLP Shared Task 2009 Workshop, pp. 10-18.

Bui, Q.C., and P.A.M. Sloot (2011), "Extracting Biological Events from Text Using Simple Syntactic Patterns," In Proceedings of BioNLP Shared Task 2011 Workshop, pp. 143-146.

Kaljurand, K., G. Schneider and F. Rinaldi (2009), "UZurich in the BioNLP 2009 Shared Task," In Proceedings of BioNLP Shared Task 2009 Workshop, pp. 28-36.

Kilicoglu, H. and S. Bergler (2011), "Adapting a General Semantic Interpretation Approach to Biological Event Extraction," In Proceedings of BioNLP Shared Task 2009 Workshop, pp. 173-182.

Kim, J.D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii (2009), "Overview of BioNLP'2009 Shared Task on Event Extraction," In Proceedings of BioNLP Shared Task 2009 Workshop, pp. 1-9.

Kim, J.D., Y. Wang and Y. Yasunori (2013), "The Genia Event Extraction Shared Task, 2013 Edition – Overview," In Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, pp. 8–15.

Kim, J.D., Y. Wang, T. Takagi and A. Yonezawa (2011), "Overview of the Genia Event task in BioNLP Shared Task 2011," In Proceedings of BioNLP Shared Task 2011 Workshop, pp. 7-15.

Miwa, M., R. Sætre, J.D. Kim and J. Tsujii (2010), "Event Extraction with Complex Event Classification Using Rich Features," In Journal of Bioinformatics and Computational Biology, vol. 8, pp. 131-146.

Nédellec, C., R. Bossy, J.D. Kim, J.J. Kim, T. Ohta, S. Pyysalo and P. Zweigenbau (2013), "Overview of BioNLP Shared Task 2013," In Proceedings of BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, pp. 1-7.

Quang, P.X., T.M. Quang and H.B. Quoc (2013), "A Hybrid Approach for Biomedical Event Extraction," in Proceedings of the BioNLP Shared Task 2013 Workshop, Association for Computational Linguistics, Sofia, Bulgaria, pp. 121–124.

Razvan, C. Bunescu, Raymond and J. Mooney (2005), "A Shortest Path Dependency Kernel for Relation Extraction," Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vacouver, B.C., pp. 724-731.

Riedel, S. and A. McCallum (2011), "Robust Biomedical Event Extraction with Dual Decomposition and Minimal Domain Adaptation," In Proceedings of BioNLP Shared Task 2011 Work-shop, pp. 46-50.

Riedel, S. and D. McClosky, M. Surdeanu, A. McCallum, C.D. Manning (2011), "Model Combination for Event Extraction in BioNLP 2011," In Proceedings of BioNLP Shared Task 2011 Workshop, pp. 51-55.

Saha, S., A. Ekbal and M. Verma (2012), "Active Learning Technique for Biomedical Named Entity Extraction," Copyright 2012 ACM 978-1-4503-1196-0/12/08.