Syntactic Dependency for Relation Extraction from Biomedical Literature

Xiaomei Wei
College of science
Huazhong Agriculture University
Wuhan, China, 027-87282492
may@mail.hzau.edu.cn

Jianyong Wang
College of science
Huazhong Agriculture University
Wuhan, China

Yang Li
College of science
Huazhong Agriculture University
Wuhan, China

***Abstract*—Relation extraction is important to improve complex natural language processing (NLP) applications. The Bio-Event extraction in GE shared task is important to understand biological processes. Although some progress has made in GE research, there is still much work to do to improve the performance of the extraction system. In this paper, we build a extraction system based on syntactic dependency technique. Relying on the output of the sentence parsing, we get rich features to build a classification model to separate the candidate edges into positive class and negative class. We obtain promising results on GE develop data set. Especially the results of simple events are comparable with the state-of-the-art GE extraction systems.**

***Keywords-syntactic parsing; Dependency; Classification; extraction*.**

I. INTRODUCTION

Event extraction is very important to improve complex natural language processing (NLP) applications[1]. In NLP, different definitions of event can be found regarding the target application. This paper is focused on the Biomedical Event Extraction which is defined in GE shared task 2011[2]. Also, as the complexity of the task was high，the gold annotation for named entities is provided. The focus is on fine-grained IE. The data for the training and development sets were derived from the publicly available event corpus [3]. We focus on the core event extraction(task 1) which addresses the extraction of typed events together with their primary arguments. Among the nine types of event, we name the first five mono-argument events as simple events while the next four as complex events.

The following sentence(given as .txt file) shows an example GE core event extraction:

(PMID1313226 abstract)*Leukotriene B4 stimulates c-fos and c-jun gene transcription.*

a set of gold protein annotation is given as .a1 file(assigned named entities an id with the prefix "T")

T1 Protein 26 31 c-fos

T2 Protein 36 41 c-jun

Our task is to extract the events from the .txt file while the named entities are provided as .a1 file. The output of the system should be .a2 file as follows:

T14 Positive_regulation 15 25 stimulates

T15 Transcription 47 60 transcription

E1 Positive_regulation:T14 Theme:E4

E2 Positive_regulation:T14 Theme:E3

E3 Transcription:T15 Theme:T1

E4 Transcription:T15 Theme:T2

This paper is organized as follows. Firstly, in Section 2 related work is reviewed. The next section provides a detailed description of our proposal to build the system. After that, Section 4 includes an evaluation of the proposal and a comparative analysis of the results. Finally, conclusions are drawn in Section 5.

II. RELATED WORK

Several approaches for extracting GEs from biomedical text have been reported. These methods range from pattern to more sophisticated machine learning (ML) systems augmented by NLP techniques such as shallow parsing or full parsing [4]. Since full parsing produces more elaborate syntactic information than shallow parsing, relation extraction systems based on full parsing can potentially obtain better results [5]. The parsed sentence can be represented either as constituent trees or dependency trees. In this case, the relation extraction task is treated as a binary classification problem

In this paper, we obtain rich features to build a classification model. Our model subsumes two tractable sub-models: one for extracting simple event edges, the other for complex event edges. We have not extracted triggers alone. Instead, in our system a joint approach is adopted to extract the event edges and identify the trigger at the same time.

III. METHODS

The workflow of the proposed system is as follows:

- Preprocessing
- Extracting biomedical event using a joint methods
- Post-processing

The overall architecture of the system is shown in Fig.1.

This research is supported by the Fundamental Research Funds for the Central Universities（Program No.2011QC063）

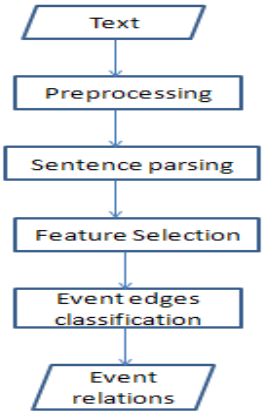

Figure 1. The overall architecture of the system

A. *Preprocessing*

1) *Sentence parsing*

Since sentence parsing produces more elaborate syntactic information than shallow parsing, relation extraction systems based on full parsing can potentially yield better results. The output of the sentence parser is shown in Table 1. Many parsers can get it. We adopt the output of McClosky parser which is provided together with the corpus.

The parsed sentence can be presented as a graph in Fig. 2.In the graph, candidate relations exist in the pairs of candidate triggers and name entities. We wi ll extract event from these pairs. From the graph, we can get the shortest paths from name entities to candidate triggers or from triggers to triggers. There are token and dependency information sequences on the paths.

TABLE I. THE OUTPUT OF SENTENCE PARSER

1 Leukotriene	_	NN	_	_	2	NMOD
2 B4	_	NN	_	_	3	VMOD
3 stimulates	_	VBZ	_	_	0	ROOT
4 c-fos	_	NN	_	_	8	NMOD
5 and	_	CC	_	_	4	COORD
6 c-jun	_	NN	_	_	5	CONJ
7 gene	_	NN	_	_	8	NMOD
8 transcription	_	NN	_	_	3	VMOD

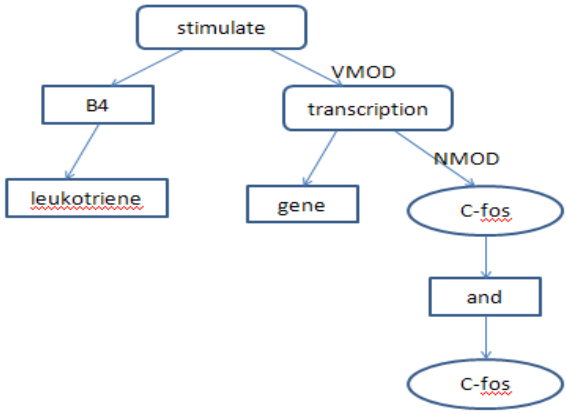

Figure 2. Typed dependency graph of the output of parser. Rounded rectangles indicate candidate triggers; ellipses mark annotated protein; rectangles mark common tokens.

2) *Acquiring candidate relation pairs*

From the ann otated data, w e build a trigger dictionary which contains stemmed trigger words and thei r frequencies that they appear as each event type.

We define the relations as a triple:

Event_type(Protein,Trigger)

Then the candidate relations can be presented as follows:

Transcription(T1,T15)

Transcription(T2,T15)

Positive _regulation(T1,T14)

Positive_regulation(T2,T14)

Positive_regulation(T15,T14)

Then the labels with prefix "T" denotes the anchor words of events and annotated name entities in a1.file.

B. *SVM classification model*

We have obtained the trigger-argument pairs and rich features of candidate relations in sec tion A. The features include Flat features and structures features. The flat features include token features and structure features denote the shortest path in the dependency path.

In allusion to the flat features and stru cture features, we employ a kernel which is a linear combination of a polynomial kernel based on flat features and Gap-weighted subsequences kernels kernel based on structure features.

The polynomial kernel is a common kernel. The experiment result[6] shows that this kernel is suitable to solve the problem like SRL(semantic role label) when d equals 2.

But polynomial kernel cannot use s tructure features which is important to our pro blem. So we in troduce Gap-weighted subsequences kernels(as described in the Fig. 3 below).

```
//Input: path sequence s and t of lengths n and m, length p, parameter λ
//Output: kernel evaluation κp (s, t) = Kern(p)
DPS (1 : n, 1 : m) = 0;
for i = 1 : n
    for j = 1 : m
```

```
        if si = tj
            DPS (i, j) =  λ2;
          end
      end
end
DP(0,0 : m) = 0;
DP(1 : n,0) = 0;
for l = 2 : p
Kern (l) = 0;
   for i = 1 : n – 1
      for j = 1 : m – 1
          DP(i, j) = DPS(i, j) +  λDP(i – 1, j)+
            λDP(i, j – 1) –  λ2 DP(i – 1, j – 1) ;
           if si = tj
              DPS (i, j) =  λ2 DP(i – 1, j – 1) ;
              Kern (l) = Kern(l) + DPS(i, j) ;
           end
       end
    end
end
```

Figure 3. Algorithm to extract candidate event pairs.

C. Post-processing

The post-processing includes three steps:

- For each trigger-protein edge of simple events, an event is generated;
- *Binding* event may have one or two theme arguments. If one edge of binding was extracted, an event is generated. IF more arguments were extracted, we will indentify the first theme argument and the second theme argument based on the syntactic rules;
- *Regulation*, *Postive_regulation* and *Negative_regulation* events may have theme arguments and cause arguments. If multiple arguments were extracted, we will indentify the "theme" argument and the "cause" argument based on the syntactic rules.

IV. EXPERIMENT AND EVALUATION

A. Datasets and evaluation

The data were prepared based on the GENIA event corpus. The data for the training and development sets were derived from the publicly available event corpus [3]. For the evaluation, we have to use the develop data as test data set since the test data evaluation online is close. We still use the evaluation standard from GE website[a]. The evaluation results are reported using the standard recall/precision/f-score metrics, under different criteria defined through the equalities.

B. Result

The final evaluation of the system was performed on the development set. Table II shows the performance of the system using *strict matching* evaluation mode defined in GE task. Our system achieved competitive results in particular some one-argument events, i.e. *Gene_expression*, *Protein catabolism, and Phosphorylation.* The result is comparable to the state-of-the-art system. But the result shows the extraction of the multiple arguments events is not perfect. Such as *Regualtion, Positive_regulation, Negative_regualtion*.

TABLE II. OUTPUT OF EVENT EXTRACTION. PER-CLASS PERFORMANCE IN TERMS OF RECALL, PRECISION, AND F-SCORE ON THE DEVELOP SET USING STRICT EVALUATION MODE.

	Gold#	**Recall**	**Prec.**	**F-score**
Gene expression	749	60.35	73.14	66.13
Transcription	158	44.94	22.4	29.9
Protein catabolism	23	95.65	73.33	83.02
Phosphorylation	111	91.89	86.44	89.08
Localization	67	67.16	78.95	72.58
=[SVT-TOTAL]=	1108	62.45	60.7	61.56
Binding	372	24.73	26.98	25.81
=[EVT-TOTAL]=	1480	52.97	52.94	52.96
Regulation	292	14.04	9.58	11.39
Positive_regulation	999	12.41	12.82	12.61
Negative_regulation	471	18.05	21.41	19.59
=[REG-TOTAL]=	1762	14.19	13.95	14.07
=[ALL-TOTAL]=	3242	31.89	31.59	31.74

C. Discussion

From the evaluation, our approach achieved competitive results in some of one-argument events. This method has some several advantages. We use a joint method of identifying the triggers and extracting event edges that is different from most of the existent system since they extract the events in a pipeline. But it is clear that the results of the multiple arguments events are not perfect. In addition to the complexity of the problem itself, we select the same features for complex events as for simple events that is proved not suitable.

V. CONSLUSION

In this paper, we propose a novel method for extracting typed events from biomedical literatures. Our approach combines the strength of both NLP techniques and ML classification. The evaluation on developed data set has shown that our system achieved results comparable to the state-of-the-art extraction methods on one-argument biomedical events. The proposed method consists of three phases: extracting candidate event pairs,partitioning data into subsets and classifying extracted candidate event pairs.

REFERENCES

[1] Sebastian Riedel, Andrew Mc Callum. Robust Biomedical Event Extraction with Dual Decomposition and Minimal Domain Adaptation. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 725–733,Beijing, August 2010.

[2] Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011b. Overview of the Genia Event task in BioNLP Shared Task 2011.

In Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task. 2011.

[3] Jin-Dong Kim, Tomoko Ohta, Jun-ichi Tsujii: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9: (2008)

[4] Björne,J. et al. 2010. Complex event extraction at PubMed scale. Bioinformatics, 26,i382–i390.

[5] Yusuke Miyao, Kenji Sagae, Rune Sætre, Takuya Matsuzaki, Jun-ichi Tsujii: Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics 25(3): 394-400 (2009)

[6] Wanxiang Che.[doctor thesis] Kernel-based Semantic Role Labeling.2008 (in Chinese)

[7] Porter, 1980, An algorithm for suffix stripping, Program, Vol. 14,no. 3, pp 130-137

[8] 8. Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: a library for support vector machines.*Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*