

Extracting Biomedical Event Using Feature Selection and Word Representation

Xinyu He, Lishuang Li*, Jieqiong Zheng, Meiyue Qin

School of Computer Science and Technology

Dalian University of Technology

116023 Dalian, China

lilishuang314@163.com hexinyu123@163.com

Abstract

We participate in the BB3 and GE4 tasks of BioNLP-ST 2016. In the BB3 task, we adopt word representation methods to improve the feature-based Biomedical Event Extraction System, and take the 4th place. In the GE4 task, based on the Uturku system, a two-stage method is proposed for trigger detection, which divides trigger detection into recognition stage and classification stage, using different features in each stage. In the edge detection, we adopt Passive-aggressive (PA) online algorithm, then we constitute events by post-processing of TEES.

1 Method

In the BB3 task, we improve the performance of the biomedical event extraction by word representation methods, which include distributed word representation, and Brown clusters representation. The framework of the proposed system includes input data, preprocessing, feature extraction, learning & classification and output data. The system preprocesses the input data from Medline literature and BB'16, and then extracts the features including word representation feature, common feature and Brown clusters feature, based on SVM classifier to learn and classify.

In the GE4 task, the system has three main components: trigger detection, edge detection and post-processing. During the trigger detection, we propose a two-stage method, which divides trigger detection into recognition stage and classification stage. During the recognition stage, we just discern the words which are trigger words, selecting the features that are more suitable for recognition; in the classification stage, we classify the triggers which are identified already, selecting the features that are more helpful to classification. In the edge detection, a multi-class PA algorithm is used, finally the events are obtained by post-processing of TEES.

	Precision	Recall	F-Score
Baseline	61.61%	38.35%	47.27%
Ours	59.91%	39.23%	47.42%

Table 1: Performance comparison on the test set

2 Experimental Results

In the BB3 task, the system achieves an F-score of 56.38% on the development set, which is 4.38 percentage points higher than the baseline. On the test set, it achieves an F-score of 47.4% on the BB3 event task, the result is shown as table 1.

In the GE4 task, the performance of our system is evaluated on the test dataset of the BioNLP'16 with online evaluation. The results related to event extraction are listed on Table 2 and Table 3.

Relations	Recall	Precision	F-Score
ThemeOf	0.51	0.50	0.51
CauseOf	0.22	0.55	0.32

Table 2: The result of relations

Denotations	Recall	Precision	F-score
Gene-expression	0.85	0.88	0.87
Binding	0.68	0.72	0.70
Localization	0.51	0.84	0.63
Phosphorylation	0.86	0.85	0.86
Potein_catabolism	0.85	0.69	0.76
ALL	0.83	0.92	0.87

Table 3: The result of denotations

3 Conclusion

In the BB3 task, our system applies distributed word representation and Brown clusters representation methods, and obtains better performance than baseline, achieving the 4th place. In the GE4 task, we adopt a two-stage method for trigger detection, which effectively avoids the situation that excessive negative samples are classified into positive samples, and the performance of the system is improved. In addition, we select different features in each stage.

Acknowledgment

This work is supported by grant from the National Natural Science Foundation of China (no. 61173101, 61173100).¹

¹*Corresponding author