**A Two-Stage Biomedical event Trigger Detection Method Integrating Feature Selection and Word Embedding**

**Abstract:**

Extracting biomedical events from biomedical literature plays an important role in the field of biomedical text mining, and the trigger detection is a key step in biomedical event extraction. We propose a two-stage method for trigger detection, which divides trigger detection into recognition stage and classification stage, and different features are selected in each stage. In the first stage, we select the features which are more suitable for recognition, and in the second stage, the features that are more helpful to classification are adopted. Furthermore, we integrate word embeddings to represent words semantically and syntactically. On the multi-level event extraction (MLEE) corpus test dataset, our method achieves an F-score of 79.75%, which outperforms the state-of-the-art systems.

**Architecture:**

****

**Introduction:**

ITH the rapid spread of Internet, the scientific biomedical literature is expanding at an exponential speed, which has made it harder than ever for scientists to research, manage, and extract knowledge from unstructured text in their research field. To tackle these problems, nature language processing (NLP) and text mining(TM) techniques are rapidly developing. In the past few years, biomedical TM has experienced the development process from biomedical named entity recognition(NER) to binary relation extraction between entities, such as drug-drug interactions(DDI) or proteinprotein interactions(PPI), and then to complex information extraction, for instance, biomedical event extraction. Biomedical event extraction focuses on the detailed behavior ofbio-molecules, which refers to the state change of one or more biomedical entities and includes expression, transcription, catabolism, phoshporylation, localization, binding and regulation of genes or proteins. As described in BioNLP’09 share task, a bio-event includes a trigger and one or more arguments, where triggers are usually verbal forms or nominalizations of verbs and may be a single word or a contiguous textual string containing multi-tokens. As shown in Fig. 1, there are three biomedical events in the sentence: the first event is a Gene expression event T1, including a trigger word "production" and a Theme type argument "1L-10"; The second event is a regulation event T2, including a trigger word "induction" and a Theme type argument T1, a Cause argument "gp41"; The last event is a Negative\_regulation event T3, including a trigger word "Prevented" and a Theme type argument T2. Event T1 is a simple event, event T2 and T3 are complex events. The structure of the three events is as follows:

Event T1 (Type: Gene\_expression, Trigger: production, Theme: 1L-10);

 Event T2 (Type: Positive\_regulation, Trigger: induction, Theme: Event T1, Cause: gp41);

Event T3 (Type: Negative\_regulation, Trigger: Prevented, Theme: Event T2).

**Existing System:**

Trigger detection is regarded as a multi-class classification task in most of the state-of-art event extractionsystems. On the commonly used dataset (MLEE), Pyysalo et al. Implemented a SVM-based approach, which manually designed salient features such as context and dependency features and fed them into a one-versus-therest SVM classifier to conduct event trigger identification for each event type, achieving an F-score of 75.84%. Zhou et al. Proposed a novel framework for event trigger identification, which learned biomedical domain knowledge from a large text corpus built from Medline and embedded it into word features using neural language modeling, achieving 78.32% F-score forevent trigger identification. However, feedforward neural language model is not optimal to train embeddings for semanticrelated tasks compared with Skip-gram model. Wang et al. Useda neural network architecture to learn better feature representation based on raw dependency-based word embeddings, and used softmax classifier to detect triggers, their methods achieved a micro F-score of 78.27% and a macro F-score of 76.94% in significant trigger classes. Proposed a word embeddings assisted neural network prediction model to conduct event identification

**Proposed System:**

Firstly, the two-stage method can alleviate the problem of class imbalance effectively. On the test dataset of the MLEE corpus, there are 14964 negative instances of all the 16720 instances, where the data imbalance is a serious problem. One stage methods classify the instances into a negative instance or the other 19 classes directly. The negative instances with large proportion may be classified incorrectly. While, our two-stage method divides trigger detection into recognition stage and classification stage. In the recognition stage, the event triggers in biomedical literatures are distinguished from non-triggers without identifying their types. And then, only the triggers which have been identified already will be classified into the correct trigger type in the second stage. In this way, it can prevent excessive negative instances from being recognized as positive instances, thus, the performance of the system will be improved. Secondly, in our two-stage method, we select different features which are more relevant for each stage to improve the performance. In the first stage, we select the features which are more suitable for recognition, and in the second stage, the features that are more helpful to classification are adopted. Finally, we integrate word embeddings to represent words semantically and syntactically in this work. Word embeddings, have achieved great success in natural language processing tasks by grouping similar words. In this paper, we utilize Skip-gram medel to trainthe word embeddings.In this paper, we present a two-stage biomedical event trigger detection approach which includes two subtasks, recognition and classification. The two stage method can alleviates the problem of class imbalance. Furthermore, different features are selected in each stage, which make the recognition and classification more effectively. Finally, word embeddings play an important role, which can learn much deeper syntactic and semantic information

**Modules:**

* **Data Preprocessing**
* **Feature Selection**
* **Feature Extraction**
* **Multi-class classification**

**Data Preprocessing:**

On the MLEE corpus, the proportion of the negative instances is far greater than the positive instances and thus we first preprocess the train set. There are two main strategies to balance negative instances and positive instances, which are Undersampling strategy and Oversampling strategy. Undersampling strategy means removing some negative instances, while, Oversampling strategy refers to increasing some positive instances. Undersampling strategy has higher F-score in the first stage than Oversampling strategy. In addition, the time cost of Undersampling strategy is less than Oversampling strategy, therefore, we utilize Undersampling strategy to balance the positive instances and negative instances. F-score of the first stage is the highest when the threshold equals 50%, therefore, we remove 50% negative instances.

**Feature Selection:**

After the preprocessing of the Turku Event Extraction System (TEES), we obtain a digital corpus, which has more than millions of feature values in each instance. In fact, a large number of features include a lot of redundant or invalid information, so it’s necessary to have feature selection to reduce the dimension of the input data, it is also helpful to shorten training time and get efficient models. We improve SVM-RFE algorithm for feature selection, as a result, the features are ranked and a valuable feature subset is obtained.

**Feature Extraction:**

The features which are suitable for multi-class classification are retained in this stage. The features are described as follows: Token features include current token text, POS, stem, binary tests for presence of uppercase, digital or special characters, bigrams and trigrams of the token. Dependency context is of great importance for trigger detection, so we extract token features of candidate triggers in dependency context and linear context besides candidate triggers themselves.

 • Token text includes current token and the tokens within a window of three tokens before and after the target tokens.

 • POS includes the POS of the current token and the tokens within a window of three tokens. The POS is tagged with McClosky-Charniak parser.

• Stem consists of the stem of the current token, obtained by Porter stemmer. This feature can alleviate the effect of morphological changes, such as “involvement” and “involves”.

**Multi-class classification:**

In the classification stage, we use PA algorithm which has good performance on our task. In order to increase the generalization and robustness of the model, it optimizes numbers of parameter C values on developing sets, and selects the mean of several outstanding classifiers’ model. It can reduce the number of combinatorial optimization parameters using the mean of models, and lower the test error rates at the time of testing.

**Algorithm:**

* **Passive-aggressive Online Algorithm:**
* **Feature-selection algorithm:**

**Passive-aggressive Online Algorithm:**

Passive-aggressive (PA) online algorithm is an online algorithm which is based on perceptron, the main idea can be summarized as follows: PA algorithm adopts maximum classification margin of SVM, using the instance greedily. Since PA algorithm follows the maximum margin theory, it has good generalization ability as SVM. It updates the classifier using the current instance greedily and predicts the current instance correctly with the maximum margin and remains the new classifier as close as possible to the current one. In order to improve the robustness of a classifier and reduce the number of possible combinations, several outstanding classifiers after optimized on the parameter C are selected and the mean of selected classifiers is adopted. In our work, the trigger multi-classification with the highest scores is regarded as the predicted results when online algorithms are used.

****

**Feature-selection algorithm:**

After the preprocessing of the Turku Event Extraction System (TEES) , we obtain a digital corpus, which has more than millions of feature values in each instance. In fact, a large number of features include a lot of redundant or invalid information, so it’s necessary to have feature selection to reduce the dimension of the input data, it is also helpful to shorten training time and get efficient models. We improve SVM-RFE algorithm for feature selection as a result, the features are ranked and a valuable feature subset is obtained. According to our experimental results, in the first stage, it achieves the best performance of 83.37%.

****

**Future work:**

Trigger detection is regarded as a multi-class classification task in most of the state-of-art event extractionsystems. On the commonly used dataset (MLEE), implemented a SVM-based approach, which manually designed salient features such as context and dependency features and fed them into a one-versus-thereSVM classifier to conduct event trigger identification for each event type, achieving an F-score of 75.84%. Zhou et al. Proposed a novel framework for event trigger identification, which learned biomedical domain knowledge from a large text corpus built from Medline and embedded it into word features using neural language modeling, achieving 78.32% F-score forevent trigger identification. However, feedforward neural language model is not optimal to train embeddings for semanticrelated tasks compared with Skip-gram model. Used a neural network architecture to learn better feature representation based on raw dependency-based word embeddings, and used softmax classifier to detect triggers, their methods achieved a micro F-score of 78.27% and a macro F-score of 76.94% in significant trigger classes. Proposed a word embeddings assisted neural network prediction model to conduct event identification. Their experimental study on the MLEE corpus achieved an F-score of 78.56. However, all the previous works for trigger detection mentioned above are mostly based on one-stage methods. There are some aspects where our two-stage method may have more advantages than the one-stage methods.

**Conclusion:**

In this paper, we present a two-stage biomedical event trigger detection approach which includes two subtasks, recognition and classification. The two stage method can alleviates the problem of class imbalance. Furthermore, different features are selected in each stage, which make the recognition and classification more effectively. Finally, word embeddings play an important role, which can learn much deeper syntactic and semantic information. By integrating the feature selection and word embeddings into the two-stage method, our system outperforms stateof-the-art performance, achieving an F-score of 79.75%. In the future, we will integrate the relation between the arguments in the same event in order to detect event edges.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL