# Using word embedding for bio-event extraction

**Chen Li**[1], **Runqing Song**[2], **Maria Liakata**[3],
**Andreas Vlachos**[4], **Stephanie Seneff**[1], **Xiangrong Zhang**[2,*]
[1] Massachusetts Institute of Technology, United States
[2] Xidian University, China
[3] University of Warwick, United Kingdom
[4] University College London, United Kingdom
[*] xrzhang@ieee.org

## Abstract

Bio-event extraction is an important phase towards the goal of extracting biological networks from the scientific literature. Recent advances in word embedding make computation of word distribution more efficient and possible. In this study, we investigate methods bringing distributional characteristics of words in the text into event extraction by using the latest word embedding methods. By using bag-of-words (BOW) features as the baseline, the result has been improved by the introduction of word-embedding features, and is comparable to the state-of-the-art solution.

## 1 Introduction

Automated extraction of bio-events from the scientific literature is an important research stage towards extraction of bio-networks, and is the main focus of bio-text-mining [1].

An event represents a biochemical process, e.g. a protein-protein interaction or chemical-protein interaction, within a signalling pathway or a metabolic pathway. An event in text is usually anchored by a word indicating the occurrence of the event, named a trigger, and the other words, which are arguments involved in the reaction. Solutions of extracting events usually begin with detecting trigger words first, and then assemble other detected argument words to a trigger. Some solutions consider event extraction as a structured prediction problem and extract triggers with corresponding arguments at once [2], [3].

BOW is common features of representing tokens when lexcial information is need for prediction, e.g. trigger prediction. However, it has drawbacks of being high dimensional, sparse and discrete. While word embedding is a collective name for a set of language modelling and feature learning techniques, by which words in a vocabulary could by mapped to vectors in a lower dimensional space, which is continuous in and relative to the vocabulary size. It is capable of representing a words distributional characteristics [4]. In this way, word embedding model may capture semantic and sequential information of a word in text. Meanwhile, a word-embedding feature is continuous, since continuous space language models maps integer vector into continuous space via learned parameters. By training a neural network language model, one obtains not just the model itself, but also the learned word embedding.

Due to the size of a dictionary word embedding might involve, computation of word distribution could be expensive. Mikolov et al. proposed two model architectures called CBOW and skip-gram for maing computation of word embedding feasible and efficient [5].

The skip-gram model tries to maximize classification of a word based on another word in the same sentence. Each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word (Figure 2).

Nie et al. utilized word embedding for detecting trigger words [6]. In this paper, we present the experiments using word embedding as token features to extract complete events including triggers and their arguments. The skip-gram model is used to obtain word-embedding features and is compared with a baseline model of using BOW features. The result demonstrates that the introduction of word embedding improves the result, and is comparable to the state-of-the-art solution.

## 2 Methods and results

### 2.1 BioNLP GENIA task

A series of efforts has been initiated to evaluate the available solutions and investigate potentials in event extraction technologies. Among them, the



*Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)*, pages 121–126,
Beijing, China, July 30, 2015. 

BioNLP Shared Tasks (BioNLP-ST) [7] have been consistently conducted since 2009 and attracted community-wide support. BioNLP-ST GENIA task is a core task and had the third edition in 2013. The task gradually increased its difficulties and complexities, for example, by upgrading from abstract-only text to full-text articles and subsuming co-reference tasks.

In the latest GENIA 2013 task, EVEX achieves the best performance (F-score: 50.97; recall: 45.44; precision: 58.03) [8]. Our system achieves a comparable result with a higher precision (F-score 47.33; recall: 37.14; precision: 65.21).

### 2.2 Event extraction model

Except binding events, the event extraction process consists of two steps in our system. First, triggers are predicted for each token in a sentence. Then, arguments including themes and causes are predicted to be associated with the triggers. The arguments could be either proteins or other events. The events, which may have other events as arguments, are called recursive events in this paper. During the prediction, this might lead to cyclic referencing. For example, event A is predicted as event Bs argument, while B is also predicted as As. In our model, the candidate events are tested, and the one has lower confidence score given by SVM classifier would be deleted. This method is also extended to bigger number of events, which are referencing each other in a cyclic manner.

For example, in Figure 1, four trigger words indicate four events. After detecting the triggers, the system check proteins one by one to seek the right arguments. The system will start with simple events, the methylation and the gene expression in the example. Then it will check arguments for the triggers of recursive events. This example has two recursive events, a positive regulation and a negative regulation. In the case when a new event is created, the new event has to be tested to see whether it could be an argument of one of the recursive events.

A binding event may have more than one theme. The extraction of binding event consists of three steps. The first two steps are similar to the other event extractions. At the third step, the candidate arguments are constructed with argument in possible combinations. Then, the combinations are tested by an SVM classifier, and the one with the highest confidence score will be kept. In the experiments, we use LibSVM as the implementation of SVM.

### 2.3 Word embedding for trigger and argument detection

Representing a token in right features is crucial in trigger prediction. BOW is a popular solution. However, it is very high dimensional, sparse and discrete. While word embedding features, which are learnt by a neural-network-based language model called continuous space language model, can represent a words distributional characteristics [4]. This, in a way, may capture semantic and sequential information of a word in text.

One problem of a word embedding model is that the model only represents the distributional characteristics of a word in entire text rather than in a specific context. In another word, the characteristics of an individual word in a sentence cannot be brought into a later prediction model. The lexically same tokens have the same word embedding. This word may indicate different event types in different sentences according to the BioNLP task. Therefore, we also experiment to join word embedding features with BOW features.

Events may have multi-token triggers. For example, mRNA expression is a transcription events trigger in many instances. Meanwhile, expression appears as a gene-expression events trigger in many instances. Biologically, transcription is a more specific process of gene expression. Therefore, for such cases, the system predicts event type as transcription since it is more informative.

In the experiment, training and development data-sets provided in the BioNLP13 are used to obtain word-embedding features in an unsupervised manner. A problem of word embedding method is that it represents a words distributional characteristics in the entire text, however loses the words contextual information in a specific sentence. Thus, during the training, we also consider $n$-gram features of a token.

After detecting triggers, assembling correct arguments to the triggers is another key link on the chain. As the model described in the section 2.1, the system starts with proteins and then the generated events. If a new event is created, it will be tested against the triggers, which indicate recursive events but have not been constructed as an event yet. The Stanford dependency path is the main feature for argument detection.

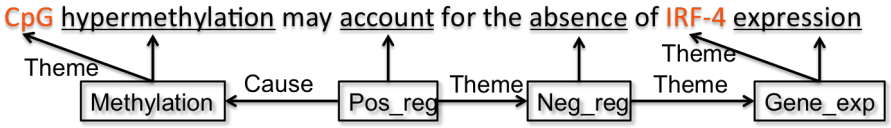


Figure 1: The model of even extraction. The words in orange are the proteins. The underlined words are the triggers.

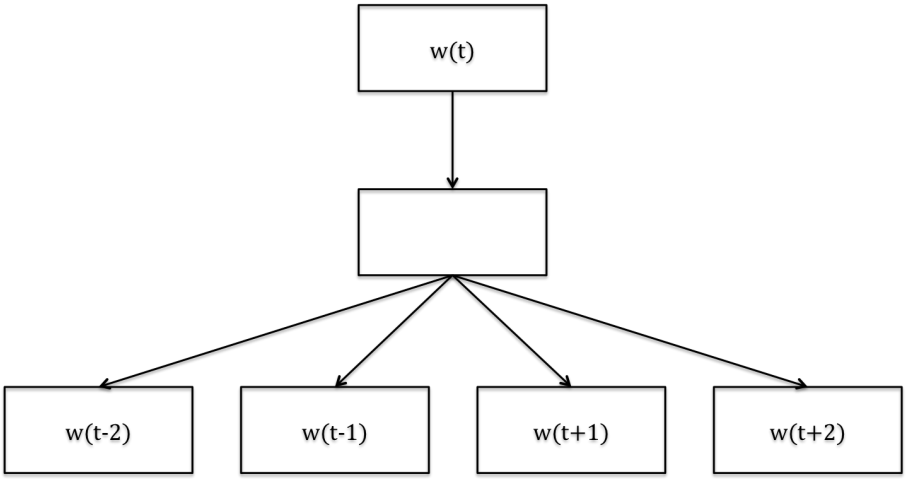


Figure 2: The skip-gram model architecture.

### 2.4 Results

We evaluate three models on the BioNLP 2013 GENIA test dataset. At the moment, only events described within the boundaries of a sentence are considered.

- BOW + *n*-gram
- Word embedding
- Word embedding + *n*-gram

The first model uses BOW and *n*-gram to represent each token. Then, the model is replaced by another using word embedding only while utilizing the exactly same extraction infrastructure, which is a pipeline converging tokenization, parsing and other pre-processing upon Apache UIMA. At last, we jointly use word embedding with *n*-gram. In Table 1, it could be observed that the joint model achieves the best performance with 47.33 in F-score. The model only using word embedding achieved the lowest, however, still gets 46.33 in F-score. This is because word embedding loses a word's distributional information in a specific context although the distributional characteristics of words are obtained for the entire text.

Table 2 shows that the detail result of the model performing the best, the joint model. Extraction of simple events achieves an average F-score of 71.98, which is expected, since each simple event contains only one theme and is not recursive. The system achieves 64.00 in F-score for protein modification event. The events are more complicated than simple events since they contain causes besides themes in arguments. The F-score for extracting binding events is 39.85. Regulatory events are the most complex ones because each of them has two arguments and is recursive. Extraction of this type of events achieved 33.97 in F-score.

Since binding is a special event type, which may have unknown number of arguments, we have analysed the extraction of binding events with different extraction strategy. Table 3 is the result with different models of assigning arguments to binding triggers. Single prediction uses one binary classifier to determine the assignment of a candidate argument. Two step prediction firstly check all arguments about whether they could be candidate arguments, then, delete the combinations covered by others. For example, if protein A and protein B are both assigned to a trigger to construct a binding event. Then, the two candidate events with A and B as argument respectively will not be considered. Two steps-confidence scores represents the results that we prune binding events ac-

| Event Class | BOW + *n*-gram | Word embedding | Word embedding + *n*-gram |
|---|---|---|---|
| Gene expression | 76.32 | 75.91 | 77.37 |
| Transcription | 59.30 | 46.39 | 60.24 |
| Protein catabolism | 64.00 | 42.55 | 64.00 |
| Localization | 51.03 | 58.39 | 45.33 |
| =[SIMPLE ALL]= | 71.66 | 68.78 | 71.98 |
| Binding | 36.36 | 35.13 | 39.85 |
| Protein modification | 0.00 | 0.00 | 0.00 |
| Phosphorylation | 72.66 | 73.68 | 70.18 |
| Ubiquitination | 12.12 | 12.12 | 12.12 |
| Acetylation | 0.00 | 0.00 | 0.00 |
| Deacetylation | 0.00 | 0.00 | 0.00 |
| =[PROT-MOD ALL]= | 66.25 | 67.46 | 64.00 |
| Regulation | 16.32 | 19.78 | 18.62 |
| Positive regulation | 36.01 | 36.74 | 35.71 |
| Negative regulation | 35.09 | 38.67 | 37.50 |
| =[REGULATION ALL]= | 33.07 | 34.45 | 33.97 |
| ==[EVENT TOTAL]== | 46.65 | 46.33 | 47.33 |

Table 1: The comparison between the BOW model, the word embedding model and the joint model on the test set of BioNLP 2013. The results are represented in F-scores.

| Event Class | Gold (match) | Answer (match) | Recall | Precision | F-score |
|---|---|---|---|---|---|
| Gene expression | 619 (441) | 521 (468) | 71.24 | 84.64 | 77.37 |
| Transcription | 101 (50) | 65 (50) | 49.50 | 76.92 | 60.24 |
| Protein catabolism | 14 (8) | 11 (8) | 57.14 | 72.73 | 64.00 |
| Localization | 99 (34) | 51 (34) | 34.34 | 66.67 | 45.33 |
| =[SIMPLE ALL]= | 833 (533) | 648 (533) | 63.99 | 82.25 | 71.98 |
| Binding | 333 (107) | 204 (107) | 32.13 | 52.45 | 39.85 |
| Protein modification | 1 (0) | 0 (0) | 0.00 | 0.00 | 0.00 |
| Phosphorylation | 160 (102) | 131 (102) | 63.75 | 77.86 | 70.10 |
| Ubiquitination | 30 (2) | 3 (2) | 6.67 | 66.67 | 12.12 |
| Acetylation | 0 (0) | 0 (0) | 0.00 | 0.00 | 0.00 |
| Deacetylation | 0 (0) | 0 (0) | 0.00 | 0.00 | 0.00 |
| =[PROT-MOD ALL]= | 191 (104) | 134 (104) | 54.45 | 77.61 | 64.00 |
| Regulation | 288 (35) | 88 (35) | 12.15 | 39.77 | 18.62 |
| Positive regulation | 1130 (291) | 500 (291) | 25.75 | 58.20 | 35.71 |
| Negative regulation | 526 (156) | 306 (156) | 29.66 | 50.98 | 37.50 |
| =[REGULATION ALL]= | 1944 (482) | 894 (482) | 24.79 | 53.91 | 33.97 |
| ==[EVENT TOTAL]== | 3301 (1226) | 1880 (1226) | 37.14 | 65.21 | 47.33 |

Table 2: The detail result on the BioNLP 2013 GENIA test dataset by using the word-embedding model.

cording to confidence scores (see the section 2.1). Table 3 shows that the performance of dividing Binding events themes extraction in two step is better. Using confidence scores to prune Binding events can improve the performance of Binding events significantly.

## 3 Conclusion

The paper explores the methods of exploiting distributional characteristics of words in a continuous space into bio-event extraction by using the latest word embedding methods. It is the first system using word embedding to extract complete events from text, and has achieve the result comparable to the state-of-the-art system's.

The system uses the BOW model as the baseline. When the model only using word embedding to represent tokens, the system achieves slightly lower performance than the BOW model's. The model jointly using word-embedding achieves the best performance. This is because *n*-gram effectively complements the loss of contextual information of words, at the same time when the words' distributional characteristics are introduced by word embedding.

There are various ways we plan to further improve the system. The current experiment uses BioNLP dataset, which is relatively small for achieving word vectors in a continuous space. In the following experiments, we would like to train and obtain the word vectors on a bigger corpus, e.g. a subset containing related articles from Wikipedia. Furthermore, we would like to create a joint model combining the prediction of trigger and arguments [3].

## Acknowledgments

The work also benefited from the discussion with Nigel Collier.

Chen Li is sponsored by Quanta Computer Inc., Taiwan.

## References

[1] C. Li, M. Liakata, and D. Rebholz-Schuhmann, "Biological network extraction from scientific literature: State of the art and challenges," *Briefings in bioinformatics*, vol. 15, no. 5, pp. 856–877, 2014.

[2] D. McClosky, S. Riedel, M. Surdeanu, A. McCallum, and C. D. Manning, "Combining joint models for biomedical event extraction," *BMC bioinformatics*, vol. 13, no. Suppl 11, S9, 2012.

[3] A. Vlachos and M. Craven, "Biomedical event extraction from abstracts and full papers using search-based structured prediction," *BMC bioinformatics*, vol. 13, no. Suppl 11, S5, 2012.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, 1988.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[6] Y. Nie, W. Rong, Y. Zhang, Y. Ouyang, and Z. Xiong, "Embedding assisted prediction architecture for event trigger identification," *Journal of bioinformatics and computational biology*, 2015.

[7] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of bionlp shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, 2013, pp. 1–7.

[8] J.-D. Kim, Y. Wang, and Y. Yasunori, "The genia event extraction shared task, 2013 edition-overview," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, 2013, pp. 8–15.

| Binding event | Gold (match) | Answer (match) | Recall | Precision | F-score |
|---|---|---|---|---|---|
| Single prediction | 333 (84) | 310 (84) | 25.23 | 27.10 | 26.13 |
| Two-step prediction | 333 (64) | 148 (64) | 19.22 | 43.24 | 26.61 |
| Two-step prediction with confidence scores | 333 (101) | 242 (101) | 30.33 | 41.74 | 35.13 |

Table 3: The results of binding event extraction on the test set of BioNLP 2013.