Biomedical Domain-oriented Word Embeddings via Small Background Texts for Biomedical Text Mining Tasks

Lishuang Li[*], Jia Wan, Degen Huang

School of Computer Science and Technology
Dalian University of Technology, Dalian China
lilishuang314@163.com

Abstract. Most word embedding methods are proposed with general purpose which take a word as a basic unit and learn embeddings by words' external contexts. However, in the field of biomedical text mining, there are many biomedical entities and syntactic chunks which can enrich the semantic meaning of word embeddings. Furthermore, large scale background texts for training word embeddings are not available in some scenarios. Therefore, we propose a novel biomedical domain-specific word embeddings model based on maximum-margin (BEMM) to train word embeddings using small set of background texts, which incorporates biomedical domain information. Experimental results show that our word embeddings overall outperform other general-purpose word embeddings on some biomedical text mining tasks.

Keywords: word embeddings; biomedical domain-oriented word embeddings; small background texts

1 Introduction

One of the most important tasks of the bioinformatics is to help biologists extract the useful data from the biomedical literature published in geometric progression. The current mainstream methods for Biomedical Information Extraction (BIE) have fully utilized word embeddings [1-2], which are usually used as extra features or inputs. Recent years have witnessed the success of word embeddings, which have been widely used in many common NLP tasks and biomedical text mining tasks, including language modeling [3], word sense disambiguation [4], semantic composition [5], entity recognition [6], syntactic parsing [7-8], biomedical named entity recognition [9] and biomedical event extraction [10]. Word embeddings have demonstrated the ability of well representing linguistic and semantic information of a text unit [11-13] and high quality embeddings can improve the performance of models. However, the word embeddings which these methods used are designed with general purpose. Experimental evidence showed that domain irrelevant word embeddings trained on large collections

[*] Corresponding author

of texts were not good enough for biomedical domain Natural Language processing (NLP) [14]. Biomedical-oriented word embeddings can outperform general-purposed ones, and further improve the performance of biomedical NLP systems. Therefore, we consider several kinds of domain-specific functional units in biomedical text, such as widely existed biomedical entities, which are incorporated in our model.

Word embeddings in previous works were often trained on large scale unlabeled texts. For example, Pennington et al. used Wikipedia, Giga word 5 and Common Crawl to learn word embeddings, each of which contained billions of tokens [15]. However, there is not always a monotonic increase in performance as the amount of background texts increase. In addition, background texts in some biomedical domain are still comparatively scarce resource such as electronic medical records, which makes it difficult to obtain large scale background texts. Thus the existing word embeddings models cannot give play to its advantages and a biomedical domain-specific word embeddings training model using small background texts is motivated. Inspired by the advantages of Support Vector Machine (SVM) in solving the small data set, we utilize maximum-margin theory and Li et al.'s method [16] to train word embeddings, which can efficiently learn high quality vector representation of words from small set of unlabeled texts.

According to the analysis above, we propose a novel biomedical domain-oriented word embeddings method based on maximum-margin, named as BEMM, which utilizes small background texts and integrates biomedical domain information. Firstly, we consider biomedical text as a sequence of words, syntactic chunks, part-of-speech (POS) tags and biomedical entities, and present a new model for the learning of word, chunk, POS and entity embeddings. Secondly, we take the advantages of SVM in solving the small data set and propose the hierarchical maximum-margin to train word embeddings on the small background texts.

The goal of the proposed word embeddings method is to improve the performance of biomedical information extraction systems such as event extraction. We compare our BEMM and the other model architectures using two different systems for biomedical text mining, i.e., the LSTM based Bacteria Biotope event extraction and Passive-aggressive (PA) [17] Online Algorithm based Biomedical Event Extraction. The experimental results show that our model has many advantages and outperforms other models on these tasks.

2 Related Work

Biomedical text mining tasks are important for the biologists. Potential information can be extracted from biomedical literatures. Utilizing the information in biomedical research contributes to the understanding of the disease mechanism and development of disease diagnosis. Word embeddings have been used in the field of the biomedical text mining tasks. Björne et al. [18] used a combination of several Long Short-Term Memory (LSTM) networks over syntactic dependency graphs for biomedical event extraction, which also added embeddings to enrich the input information. Li et al. [19] utilized a hybrid method that combined both a rule-based method and a machine learning-based method to extract biomedical semantic relations from texts, which used word embeddings as features.

Representation of words as continuous vectors has a long history [20-21]. There are two ways to learn the word embeddings.

Matrix factorization methods for generating low-dimensional word representations have roots stretching as far back as LSA. These methods utilize low-rank approximations to decompose large matrices that capture statistical information about a corpus. The latest efficient word representation model based on matrix factorization is GloVe, by which Pennington et al. [15] leveraged statistical information by training only on the nonzero elements in a word-word co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus.

Another approach is to learn word representations that aid in making predictions within local context windows. Neural network structure as a popular model architecture has also played an important role in learning useful word embeddings. For example, Bengio et al. [3] proposed neural network language model (NNLM), where a feedforward neural network with a linear projection layer and a non-linear hidden layer was used to learn the word vector representation and a statistical language model. Recently, Mikolov et al. [11] proposed two efficient models, continuous bag-of-words model (CBOW) and skip-gram model, to learn word embeddings from large-scale text corpora. The training objective of CBOW is to combine the embeddings of context words to predict the target word, while skip-gram is to use the embedding of each target word to predict its context words. These two architectures both greatly reduce the computational complexity. In this work, we directly extend these architectures, and focus on the training of word embeddings suitable for the field of biomedical words.

3 Method

In this section, we explore a biomedical domain-specific word embeddings model based on maximum-margin to train word embeddings using small set of background texts, which incorporates domain information such as biomedical entities. Figure 1 shows the overall framework of our method, which consists of two components, extracting functional units (Section 3.1) and training of embeddings based on maximum-margin integrating functional units (Section 3.2). In the following sections we will illustrate our method in detail.

3.1 Extracting functional units

In the stage of extracting functional units, our main work is to obtain the background texts and process it. Firstly, we download abstract texts from the PubMed to get the background texts. Then the texts are split into sentences, and we tokenize the sentence into atomic units. Finally, we use GENIA Dependency parser (GDep) [22] to extract functional units containing biomedical information. GDep is often used to parse the sentences in the biomedical domain, so we utilize it to process our background texts.

The functional units we extract include stem, chunk, entity and POS tags, which are described in detail below. Our model contains biomedical domain information compared to the skip-gram model only using the information of the words itself. And we take the input sentence "contains pathogenic bacteria in red blood cells" as an example.

Stem. Although we all know that "contains" is the third person singular form of "contain", the machines do not realize it. Instead, the machines regard "contain" and "contains" as two totally different tokens. Considering the stem can solve the problem.

POS tags. POS reflects the role of words in a sentence, which is important for the analysis of a sentence structure. So we use the POS to train embeddings to gain more information.

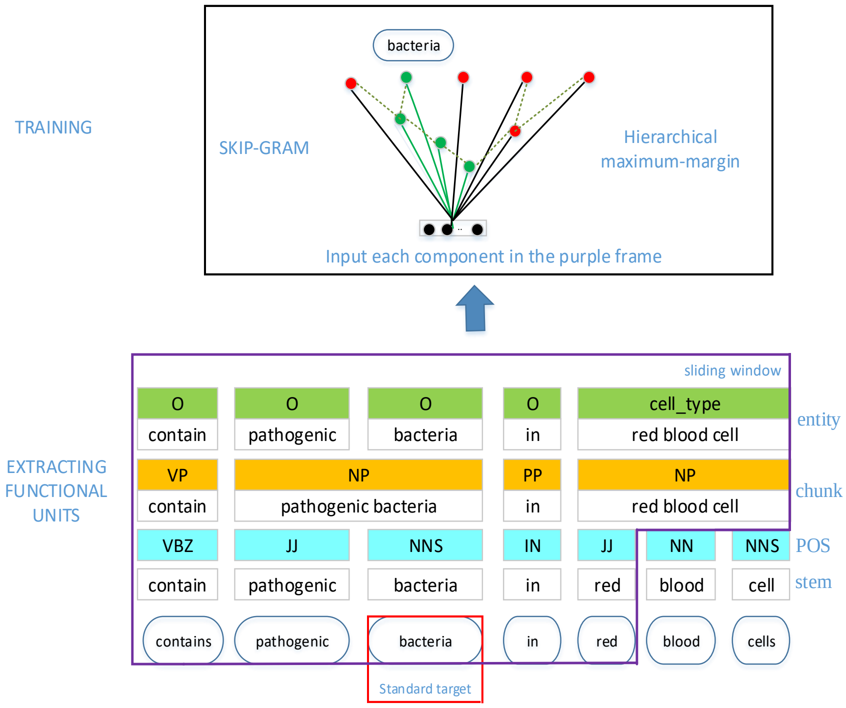

Fig. 1. The overall framework of our method.

Chunk. Only considering words and stems may be not enough, because in this case the model trains “pathogenic” using “bacteria”, ignoring the truth that “pathogenic” is the modifiers of “bacteria”. However, “pathogenic bacteria” should be considered as a whole. Therefore, in our method, we consider syntactic chunks as a parallel layer of word sequences.

Entity. While the chunk helps understanding the meaning of words, biomedical entities can improve the word embeddings from another perspective. Entity layer provides fine-grained understanding of biomedical text.

After getting all functional units from the texts, we utilize sliding window to obtain the context information of the target word, including the words and functional units, by which the target word would be trained. Many neural embeddings use a sliding window of size *k* around the target word *w*, so 2*k* contexts are produced: the *k* words before *w* and the *k* words after *w*. For *k* = 2 in Figure 1, the contexts of the target word *w* are w_{-2}, w_{-1}, w_{+1}, w_{+2} and their functional units.

3.2 Training based on maximum-margin integrating functional units

As Figure 1 shows, all the functional units in the purple box, including words, stems, chunks, POS and entities, are trained using the standard target word in the red box. In order to minimize the computational complexity, our model architecture removes the hidden layer. In addition, the training model uses a hierarchical maximum-margin strategy, which firstly sorts all the units by frequency of words in the vocabulary, then makes the two lowest frequency elements into leaves and creates a parent node with a frequency that is the sum of the two units' frequencies, and finally repeats this step until all units are included in a Huffman tree. All units are leaf nodes, and the units with lower frequency have higher depths and longer binary codes.

For each target word, the corresponding word vector is trained using the back propagation algorithm. For a given surrounding word *y*, the objective of the word embeddings is to let the prediction $\hat{y}^{(\mathrm{i})}$ equal to $y^{(\mathrm{i})}$, where $x^{(i)}$ is the corresponding word vector of target word w_i, and θ is the parameter of the model (weight matrix). t is the label of each category. We introduce the idea of maximum-margin. The cost function is with equation (1).

$$J(\theta) = \max\left(0, 1 + \max_{t \neq y^{(i)}} \theta_t^{\mathrm{T}} x^{(i)} - \theta_{y^{(i)}}^{\mathrm{T}} x^{(i)}\right). \tag{1}$$

On each iteration if the equation (2) is satisfied:

$$1 + \max_{t \neq y^{(i)}} \theta_t^{\mathrm{T}} x^{(i)} - \theta_{y^{(i)}}^{\mathrm{T}} x^{(i)} > 0\,, \tag{2}$$

We perform with equation (3) where α is the learning rate:

$$\theta_t \coloneqq \theta_t - \alpha x^{(i)}\ , \tag{3}$$

and equation (4):

$$\theta_{y^{(i)}} \coloneqq \theta_{y^{(i)}} + \alpha x^{(i)}\,, \tag{4}$$

and equation (5):

$$x^{(i)} \coloneqq x^{(i)} - \alpha\left(\theta_t - \theta_{y^{(i)}}\right). \tag{5}$$

3.3 Complexity of the model

Our proposed model architecture is similar to the skip-gram model. We use each target word as an input to a log-linear classifier with continuous projection layer, and predict functional units within a certain window before and after the target word. In our model architecture, the non-linear hidden layer is removed and a log-linear classifier is built. For the model, the training complexity is proportional to the equation (6):

$$\mathrm{O} = \mathrm{E} \times \mathrm{T} \times \mathrm{Q}\,, \tag{6}$$

where *E* is the number of the training epochs and *T* is the number of the words in the background texts. *Q* can be represented as the equation (7):

$$\mathrm{Q} = \mathrm{C}\ \times (\mathrm{D} + \mathrm{D} \times\ \log_2 \mathrm{V}), \tag{7}$$

where *C* is the functional units before and after the target words, *D* is the dimension of the input layer and *V* is size of the vocabulary.

4 Experiments

The objective of this study is to train word embeddings which represent biomedical semantic regularities and further improve the performance of BIE systems. To evaluate the performance of the word embeddings, we download background documents from PubMed to train word embeddings, and then apply them into two different BIE systems, respectively for Bacteria Biotope (BB) event extraction [1] and Biomedical Event Extraction (BEE). To make a careful comparison between the widely used other model architectures and our model, we design two different experiments. One is that both models utilize the same small background texts to train word embeddings for BB event extraction. The other is that the other model architectures employ the large scale background texts and our model applies the small scale background texts for BEE.

Note that 1) the two biomedical text mining tasks are only used for evaluating the word embeddings, so we just explain the tasks briefly in the following sections, and 2) since the two systems are not the main focus of this paper, we only introduce the general schemes of them, and 3) we train word embeddings using skip-gram and our BEMM with same parameters, e.g., same size of sliding window 5, same starting learning rate 0.025, and same word vector dimension 50, and 4) we aim to obtain state-of-the-art word embeddings, not biomedical text mining systems, therefore, we focus on the comparison of word embeddings rather than biomedical systems.

4.1 Bacteria Biotope event extraction

The purpose of the BB task is to study the interaction mechanisms of the bacteria with their environment from genetic, phylogenetic and ecology perspectives. The BB event extraction task has been put forward in the BioNLP-ST[1].

The BB event extraction can be treated as a classification problem. We use LSTM with a top layer of softmax to classify each BB instance and word embeddings are used as inputs in the LSTM model. In this paper, we evaluate six word embeddings models respectively on different scale corpora leveraging BB event extraction task. These six models are presented below: the skip-gram model and the cbow model in the Word2Vec tool; the model of glove; the model based on functional units not using maximum-margin theory, named as BE; the model based on maximum-margin theory not using functional units, named as MM; the model integrating functional units and maximum-margin theory, i.e. BEMM.

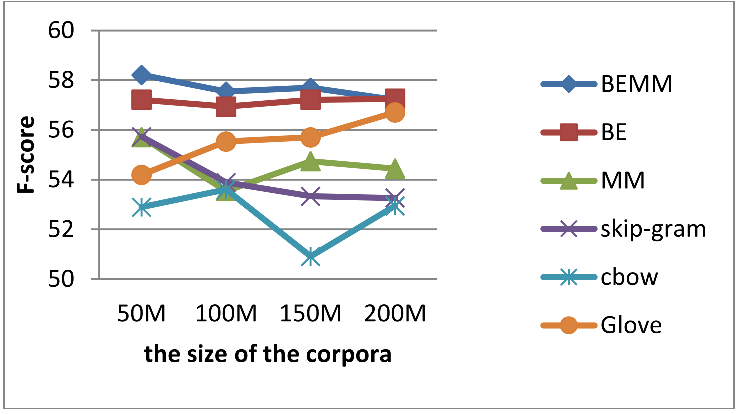

Fig. 2. The results of six models for the BB event task.

The corpora are the abstracts downloaded from PubMed, which use "bacteria" as the key word. The sizes of the corpora after the GDep parser are 50M, 100M, 150M, and 200M. We use a window size of 5 and the threshold for sub-sampling is set to 0; the initial learning rate is set to 0.025; and we keep all words that occur in the background text. We train word embeddings with 50 dimensions and also try different dimensions such as 100 dimensions, 200 dimensions and 400 dimensions whose F-score is 56.89%, 56.25% and 56.77% respectively. So we only show the result of 50 dimensions.

The results are shown in Figure 2. We could observe when we use the small background texts to train word embeddings, our BEMM, BE models significantly outperform other models. The rationality is analyzed as follows: 1) The word embeddings trained by BE contain more domain information, so the LSTM can learn more from the inputs when the word embeddings are used as inputs. 2) Maximum-margin theory has a relatively good performance in the small background texts, which has been veri-

[1] http://2016.bionlp-st.org/tasks/bb2

fied on the SVM. Therefore, the word embeddings trained by MM could perform well. 3) The model BEMM combining functional units and the maximum-margin has gained both strengths. It can not only get more semantic information, but also be able to be competent in small background texts.

Table 1. Comparison with existing system

Methods	F-score	Recall	Precision
System using BEMM	**58.21%**	**66.96%**	51.47%
Li's method	57.14%	57.99%	**56.32%**

We also compare the performance between the system using BEMM and the currently first-rate system [23] for the BB event task. The results are shown in Table 1.

Both methods apply LSTM to conduct the BB event task using word embeddings. From the Table 1, we can see that the F-score of system using BEMM is 1.07 percentage points higher than the currently best system, which proves that our BEMM has a positive effect on the results.

4.2 Biomedical Event Extraction

To further evaluate the generalization ability of our model, we have also experimented with another BIE task, namely biomedical event extraction. Biomedical event extraction focuses on the detailed behavior of bio-molecules, which refers to the state change of one or more biomedical entities and includes expression, transcription, etc. Our experiments are conducted on the commonly used dataset MLEE [24].

We compare two word embeddings based on the same PA model. One is the word embeddings trained by our model BEMM on small background texts, and the other is the word embeddings trained by the skip-gram model on large background texts. Parameter settings are described below.

- *Word Embeddings on Small Background Texts.* The background texts are abstracts downloaded from PubMed with different key word "protein". The dimensions of word embeddings are set to 400; the size of the background text after the GDep parser is 50M. And the rest parameters are set the same as the previous experiment.
- *Word Embeddings on Large Background Texts.* The way we train word embeddings is as follows: First, abstract texts are downloaded from the public database, PubMed, with the size of about 5.6G. Then, all abstracts are split into sentences and tokenized into tokens. Finally, all tokenized sentences are sent into the skip-gram model. The parameters, window sizes and dimensions, are set 7 and 400 respectively.

Li et al.'s [25] have demonstrated the effectiveness of the word embeddings of 400 dimensions on BEE task. Therefore, we use the embeddings of 400 dimensions. To verify the effect of word embeddings on the performance of the task, we set all the parameters to the same in the PA model except for the word embeddings and the results are shown in Table 2.

Table 2. Comparison between BEMM and the skip-gram for BEE task.

Word embeddings	F-score	Recall	Precision
BEMM	**79.37%**	**78.01%**	**80.77%**
skip-gram	76.54%	73.29%	80.09%

From Table 2, we can observe that the F-score of our method is 2.83 percentage points higher than the other, because our model makes full use of the biomedical semantic information, although it only uses small scale background texts. The experimental results prove that our approach is effective on small background texts.

5 Discussion

From the above experimental results, we can conclude that our model outperforms well on the biomedical text mining tasks and mainly includes the following important advantages:

Integrating Functional Units. We do not just use the word itself to obtain the word embeddings, but also fully exploit linguistic and biomedical functional units such as entities, which are integrated into word embeddings. Therefore, our model can achieve the vector representation of each word and functional units, and we apply these word embeddings with rich semantic information to the biomedical text mining tasks, so that the model can learn more and the results are more accurate.

Utilizing Maximum-margin Theory. The maximum-margin theory can well handle small training samples. Therefore, the maximum-margin classification is used to train our model in order to solve the problem of insufficient corpus, which makes the word embeddings trained by small corpus still has a good performance. In our framework, at the output layer, we use the hierarchical structure which is implemented by the Huffman tree. Then, we take advantage of the maximum-margin classification to make predictions in each branch of the tree.

Temperate Computational Complexity. Our framework structure is similar to the skip-gram model, so our model also has a low computational complexity. Under the same corpus and the number of the iterations, the computational complexity of our model and the skip-gram model are of an order of magnitude. Overall, the computational complexity of our model is moderate.

6 Conclusion

This paper proposes a novel biomedical domain-oriented word embeddings model using small background texts which integrates the biomedical semantic information. There are three major contributions of this work:

Firstly, we incorporate functional units including stem, POS tags, chunk and entities into the word embeddings to enrich semantic expression.

Secondly, we utilize the maximum-margin theory to deal with the problem of insufficient background texts.

Thirdly, the experimental results show that our word embeddings are effective on BIE applications.

In the future works, we plan to consider more biomedical domain knowledge to train better word embeddings for biomedical text mining applications, for example, deep syntactic information, external knowledge bases such as Gene Ontology, Drug Bank, etc.

Acknowledgment. The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under NO. 61672126.

References

1. Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferré, A., Bessières, P.: Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016. In: Bionlp Shared Task Workshop - Association for Computational Linguistics, pp. 12-22 (2016)
2. Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M.: Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. In: Bionlp Shared Task Workshop - Association for Computational Linguistics, pp.1-11 (2017)
3. Bengio, Y., Vincent, P., Janvin, C. A neural probabilistic language model. Journal of Machine Learning Research, 3(6), 1137-1155 (2003)
4. Chen, X., Liu, Z., Sun, M.: A Unified Model for Word Sense Representation and Disambiguation. In: Conference on Empirical Methods in Natural Language Processing, pp.1025-1035 (2014)
5. Zhao, Y., Liu, Z., Sun, M.: Phrase type sensitive tensor indexing model for semantic composition. In: Twenty-Ninth AAAI Conference on Artificial Intelligence, pp.2195-2201 (2015)
6. Collobert, R., Weston, J., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(1), 2493-2537 (2011)
7. Socher, R., Lin, C. Y., Ng, A. Y., Manning, C. D.: Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In: International Conference on Machine Learning, ICML 2011, pp.129-136 (2011)
8. Socher, R., Bauer, J., Manning, C. D., Ng, A. Y.: Parsing with Compositional Vector Grammars. In: Meeting of the Association for Computational Linguistics, pp.455-465 (2013)
9. Tang, B., Cao, H., Wang, X., Chen, Q., Xu, H.: Evaluating word representation features in biomedical named entity recognition tasks. Biomed Research International, 2014(2), 1-6 (2014)
10. Li, C., Rao, Z., Zhang, X.: LitWay, Discriminative Extraction for Different Bio-Events. In: Bionlp Shared Task Workshop, pp.32-41 (2016)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computer Science. (2013)
12. Mikolov, T., Yih, W. T., Zweig, G.: Linguistic regularities in continuous space word representations. In: In Proceedings of the Conference of the North American Chapter of the

Association for Computational Linguistics: Hu-man Language Technologies, pp. 746–751 (2013)

13. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In: Meeting of the Association for Computational Linguistics, pp.1555-1565 (2014)
14. Jiang, Z., Li, L., Huang, D., Jin, L.: Training word embeddings for deep learning in biomedical text mining tasks. In: IEEE International Conference on Bioinformatics and Biomedicine, pp.625-628 (2015)
15. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Conference on Empirical Methods in Natural Language Processing, pp.1532-1543 (2014)
16. Li, L., Jiang, Z., Liu, Y., Huang, D.: Word Representation on Small Background Texts. In： Social Media Processing, pp.143-150 (2016)
17. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research, 7(3), 551-585 (2006)
18. Mehryary, F., Björne, J., Pyysalo, S., Salakoski, T., Ginter, F.: Deep Learning with Minimal Training Data: TurkuNLP Entry in the BioNLP Shared Task 2016. In: Bionlp Shared Task Workshop, pp.73-81 (2016)
19. Li, L., Qin, M., Huang, D.: Biomedical Event Trigger Detection Based on Hybrid Methods Integrating Word Embeddings. In: Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data, pp.67-79 (2016)
20. Hinton, G. E., McClelland, J., Rumelhart, D. E.: Distributed representations. Parallel distributed processing: Explorations in the microstructure of cognition, (1), 77-109 (1986)
21. Rumelhart, D. E., Hinton, G. E., Williams, R. J.: Learning representations by back-propagating errors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 323(6088), 533-536 (1986)
22. Sagae, K., Tsujii, J. I.: Dependency Parsing and Domain Adaptation with LR Models and-Parser Ensembles. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp.1044–1050 (2007)
23. Li, L., Zheng, J., Wan, J., Huang, D., Lin, X.: Biomedical event extraction via Long Short Term Memory networks along dynamic extended tree. In: IEEE International Conference on Bioinformatics and Biomedicine, pp.739-742 (2016)
24. Pyysalo, S., Ohta, T., Miwa, M., Cho, H. C., Tsujii, J., Ananiadou, S.: Event extraction across multiple levels of biological organization. Bioinformatics, 28(18), 575-581 (2012)
25. Li L., Liu S., Qin M., Wang Y., Huang D.: Extracting Biomedical Event with Dual Decomposition Integrating Word Embeddings. Transactions on Computational Biology and Bioinformatics, 13, 669-677 (2015)