Journal of Information Hiding and Multimedia Signal Processing ©2020 ISSN 2073-4212
Ubiquitous International Volume **11**, Number **2**, June **2020**

Extraction of English Drug Names Based on Bert-CNN Model

Yu Zhang

Dalian Polytechnic University, Dalian, China
18840848215@163.com

Li Guo and Chenning Du

Dalian Polytechnic University, Dalian, China
guoli@dlpu.edu.cn; duchenning@163.com

Yu Wang and Degen Huang

Dalian Polytechnic University, Dalian, China;
Dalian University of Technology, Dalian, China
wy18941178080@163.com; huangdg@dlut.edu.cn

Received April 2020; revised June 2020

Abstract. *Drugs bring good news to people's health, while the adverse reactions caused by taking two or more drugs at the same time will aggravate the damage to patients' health. Therefore, the research on the interaction between drugs has been a hot topic in the biomedical field. This paper proposes the method of English drug name relation extraction based on the Bert-CNN. Usually, the methods of the extraction of entity relationship are mostly based on the word vector trained by Word2vec, Glove, and so forth. The problem of those methods is that they can't distinguish the different semantics of polysemous words. In this paper, Bert is used to train the word vectors. The word vectors generated by Bert are dynamically represented by the surrounding words of the word. Then, word vector is used as the high-quality feature input of the downstream CNN. F1 value of the proposed method obtained on DDIExtraction2013 dataset is 72.64%.*
Keywords: Relation Extraction, Bert, CNN

1. **Introduction.** With the development of medicine and the deeply researching on DDI, a lot of valuable medical information is hidden in the structured and unstructured medical literature, which is growing exponentially. In order to find important information from biomedical literature, information extraction technology has attracted much attention in recent years. At present, information extraction in biomedical literature is mainly entity extraction and relationship extraction, however, this paper only focuses on relationship extraction.

In the existing research, there are two kinds of methods about drug name extraction task: rule-based method and machine learning method. It is necessary for the rule-based method to analyze the relationship between texts and label first and then summarize the rule model. For example, Segura-bedmar et al.[1] extract the rule set in the corpus, and use the rule set to find the drug pairs matching the rules in the corpus. Although the accuracy of those methods is high, the recall rate is poor, and the F value is reported only about 19%. The performance of using rule-based method to extract model performance

relies on the ability of set extraction rules of professionals and domain experts. Therefore, machine learning based method for drug name extraction has been a trend.

The method based on machine learning usually transforms the extraction of drug name relationship into a classification task. Wang[2] proposed a support vector machine (SVM) method based on multiple features, including: word features, location features, negative word features and sentence distance features. Kim[3] also input rich features into SVM model for drug name relationship extraction. The main features used in this method are: word features, syntax tree feature core and noun phrase constraint cooperative features, etc. and this method has achieved 67% F value in DDIExtraction2013 dataset.

The above methods need to extract features manually, but it is time-consuming and subjective. Recently, the convolutional neural networks (CNNs) as one of the deep learning methods have made great achievements in sequence tagging[4], emotional analysis[5], etc. Based on this, we use CNN to extract the drug name relationship.

DDIExtraction2013 dataset is composed of multiple sentences. In this paper, based on sentence level, dynamic word vector and position information generated by Bert are used as the input vectors of CNN to extract the relationship between drug name pairs. Experimental results show that the performance of drug name relation extraction based on Bert-CNN is better than that based on SVM.

2. **RELEVANT WORK.**

This section introduces the extraction model of English drug name relationship based on Bert-CNN. Figure 1 shows the four layer architecture of the model to extract drug name entity relationship: embedding layer, convolution layer, pooling layer and softmax layer.

2.1. **Word Coding Layer.**

Word vector is a way to digitize words in spoken language. In natural language processing tasks, word vectors have two forms: one hot representation and distribution representation. Compared with the discrete representation, the distributed representation has the following two advantages: (1) there is similarity between words; (2) the word vector can contain more information, and each dimension has a specific meaning. The task of NLP using deep learning usually uses a distributed representation, which represents the word as a continuous dense vector of fixed length. At present, there are many models that can be used to learn word vectors, such as Word2vec[6] and Golve[7]. But distributed representation can't solve the problem of polysemy. In natural language, every word may have many different meanings. If we use numerical value to represent its meaning, at least it should not be a fixed vector.

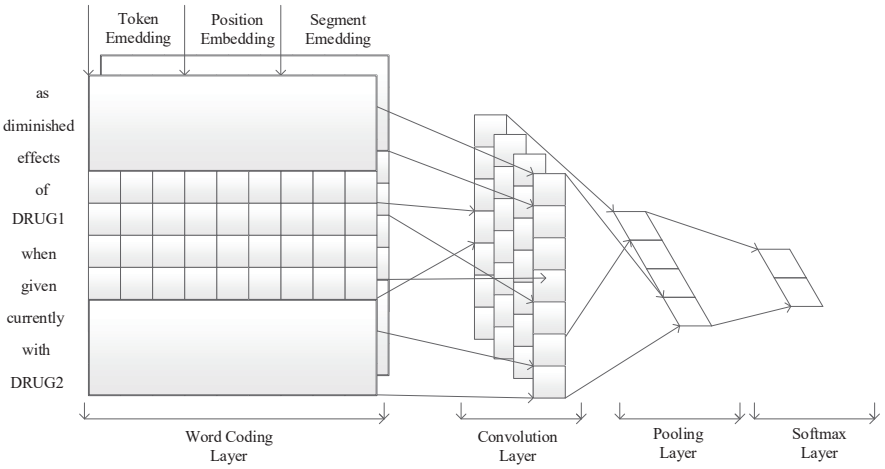

FIGURE 1. Model architecture

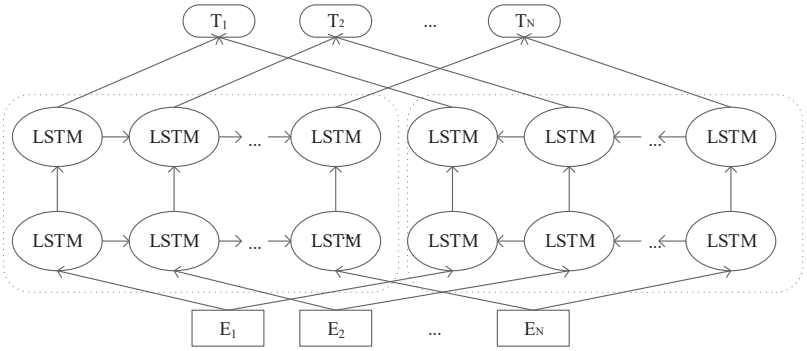

FIGURE 2. ELMo model

Peters, et al.[8] proposed embedding from language models (Elmo). Elmo applied two stacked long short term memory (LSTM) layers to learn context information for each word in two directions of a sentence. The final vector representation of each word is composed of hidden states corresponding to several layers of LSTM, as shown in Figure 2.

In addition, Radford[9] and others put forward the OpenAI GPT model, which is based on the language model of transformer, using the structure of transformer to train the one-way language model, and then the downstream natural language processing tasks can be fine-tuned on this basis. Compared with LSTM, GPT has the advantage that it can get the language information of the sentence context in a long distance. The GPT model is shown in Figure 3.

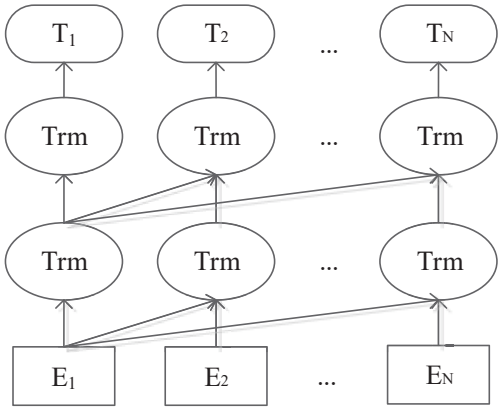

FIGURE 3. GPT Model

In order to make full use of the context information of sentences, Devlin[10] and others proposed the Bert model, as shown in Figure 4. Bert uses two-way transformer, which combines the advantages of the above models and removes their disadvantages. Its feature representation is the context from left and right sides depended by all layers. It has achieved good results in many natural language processing specific follow-up tasks. In the word coding layer of this paper, we can directly use Bert's feature representation as the word embedding feature. The coding vector of Bert input is the sum of three embedding features: token embedding, position embedding and segmentation embedding.

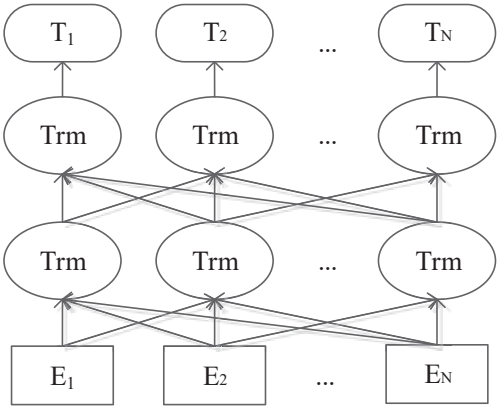

FIGURE 4. Bert Model

2.2. **Attention Layer.** Convolution layer realizes CNN's automatic feature extraction function, which can extract the features of the input data no matter whether it is one-dimensional, two-dimensional or multi-dimensional. Each convolution layer can have multiple convolution kernels, and different convolution kernels have different extraction characteristics. In this paper, convolution kernels with heights of 2, 3, 4 and 5 are used. The convolution kernel is the size of the receptive field between the neurons in the convolution layer. The convolution layer can realize weight sharing, which can greatly reduce the training time of the network compared with the fully connected neural network. The weight can be updated by error back propagation.

The features extracted by convolution layers are linear, which is followed by activation function to make it nonlinear to store more information and enhance the ability of feature expression. The commonly used activation functions are tanh[11], sigmoid[12] and relu[13].

Assuming that the height of the convolution kernel is m, the characteristic hi can be calculated by the continuous m word vectors in the convolution kernel and sentence vector, and the calculation formula is shown in Formula 1.

$$h_i = \tanh\left(w^i \cdot X + b\right) \tag{1}$$

Where i represents the $i-th$ word, w_i represents the weight of the $i-th$ word, and X represents the continuous word vector, with $X = [x_i, \ldots, x_{i+m-1}]$, b is the offset matrix. So every word $i(i \sim [1, n])$ in a sentence can be expressed as $H = [h_1, \ldots, h_{n-m+1}]$, its length is $n - m + 1$

2.3. **Pooling layer.** This layer compresses the input feature map information, which can greatly reduce the feature dimension, thus reducing the risk of over fitting. Common pooling operations include mean pooling and Max pooling, as shown in Figure 5. In this paper, we use the maximum pooling operation, which is to select the point with the maximum value in the local acceptance domain.

2.4. **Sotfmax layer.** The final features from the pool layer are input into the sotfmax layer for classification. In the process of training, in order to avoid the phenomenon of data over fitting, dropout technology is used in this paper, that is to say, the pooled vector sets the feature to 0 with a certain probability, instead of completely sending it to the softmax layer for classification. This operation is not required for testing.

3. EXPERIMENT AND ANALYSIS.

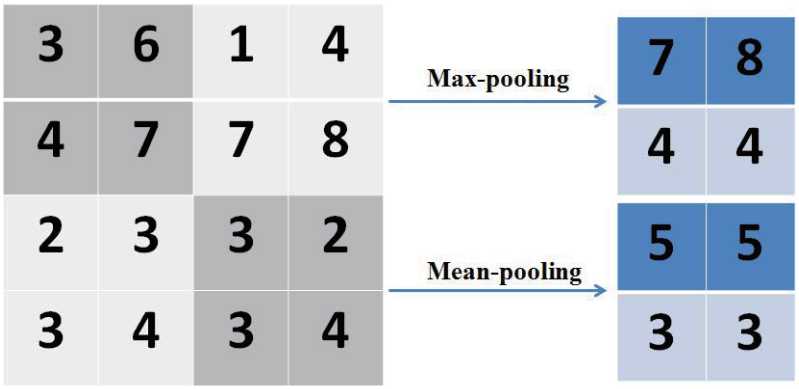

FIGURE 5. Pooling Layer Operation

3.1. **DataSet.** The experimental data used in this paper is the DDIExtraction2013 evaluation data set published by Segura-bedmar et al. The data set mainly comes from the Drugbank database and MEDLINE database. The classification of drug name relationship in the data set is divided into the following four categories: (1) advice class: the description of drug interaction in the text is a suggestion or recommendation. (2) effect class: the text describes that the drug interaction will have some influence or have some pharmacodynamic mechanism. (3) mechanism class: the text describes drug interactions based on the mechanism of pharmacokinetics. (4) int class: the text describes the drug interaction between drugs, but there is no other description.

The specific statistics of the experimental data set are shown in Table 1. Among them, positive is the statistics of all positive cases, i.e., there is a relationship between drug name entities, while, negative is the statistics of all negative cases, which means that there is no relationship between drug name entities.

TABLE 1. EXPERIMENTAL DATASET

	Training Set			Test Set		
	DrugBank	MedLine	ALL	DrugBank	MedLine	ALL
Negative	22118	1547	23772	4367	345	4712
Positive	3788	232	4020	884	95	979
advice	818	8	826	214	7	221
effect	1535	152	1687	298	62	360
mechanism	1257	62	1319	278	24	302
int	178	10	188	94	2	96

In DDIExtraction2013 dataset, there are about 10% positive cases and 90% negative cases. The distribution of positive cases and negative cases is seriously uneven, which will interfere with the actual extraction of drug name relationship. In order to alleviate the imbalance between positive and negative cases, according to the method proposed in [14, 15], this paper filters out negative cases as much as possible in the following two cases.

If the selected pair of drug name relationship has the same name, delete the corresponding instance, or the format of two drug names is different, but the same drug. Because a drug can't have any relationship with itself. For example: Animal toxicology studies showed increased DEET[DRUG1] toxicity when DEET[DRUG2] was included as proof of the formulation.

If two or more drug names are juxtaposed and two drug names are grammatically juxtaposed, then there will be no general relationship. For example: Before using this medication, tell your doctor or pharmacist of all prescription and nonprescription products you may use, especially of: amino-glycosides (e.g., gentamicin[DRUG1], amika-cin[DRUG2]), amphotericin B, cyclosporine, non-steroidal anti-inflammatory drugs (e.g. ibu-profen), tacrolimus, vancomycin.

The processed data set statistics are shown in Table 2. In the experiment, three evaluation indexes, accuracy P, recall R and F1 value, were used to evaluate the results of drug name relationship extraction. Among them, F1 is the performance of the comprehensive evaluation model. The three evaluation indexes are shown in formula 2-4:

$$P = \frac{TP}{FP + TP} \tag{2}$$

$$R = \frac{TP}{FN + TP} \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Table 2. DATASET AFTER PROCESSING

	Training Set			Test Set		
	DrugBank	MedLine	ALL	DrugBank	MedLine	ALL
Negative	17571	1249	18820	3426	310	3736
Positive	3775	231	4006	884	90	974
advice	816	7	823	214	7	221
effect	1531	152	1683	298	57	355
mechanism	1254	62	1316	278	24	302
int	174	10	184	94	2	96

Among them, true positive (TP) is a positive case with correct classification, false positive (FP) is a positive case with wrong classification, and false negative (FN) is a negative case with wrong classification.

3.2. **Experiment parameter setting.** In this experiment, the open-source Python framework of Facebook artificial intelligence research institute is used. In this paper, we first apply Bert training word vector to the extraction of English drug name relationship based on Bert CNN. Because CNN model needs fixed input text length, and the actual length of each text is not the same, so we set a unified sentence length of 150. The experimental parameters used in this paper are shown in Table 3.

Table 3. DATASET AFTER PROCESSING

Parameter names	The parameter value
Batch number	16
Word vector dimension	768
Convolution window	2,3,4,5
Iteration times	100
Optimization algorithm	dam
Maximum sentence length	150
Dropout rate	0.5

3.3. **Experimental results and analysis.** Table 4 shows the relationship extraction and detailed evaluation of all categories of Bert-CNN model used in this paper. According to Table 4, it can be observed that this method can correctly classify advice, effect and mechanism, and F1 value is greater than 70%. There are some difficulties in int classification, because the low proportion (¡2%) of int class in training data set leads to the neglect of the difference between this class and other classes in feature extraction, so it can not be effectively identified.

Table 4. EXPERIMENTAL RESULTS OF DDL2013

Results		P	R	F1
category	advise	77.87	82.81	80.26
	effect	66.84	73.8	70.15
	mechanism	82.03	69.54	75.27
	int	85	35.42	50
Micro-Average		74.65	70.74	72.64

In order to further evaluate the performance of the system, this paper lists the results of various other model experiments on the DDIExtraction2013 data set and compares them with the Micro-Average results of this experiment. The results are shown in Table 5.

In Table 5, the methods proposed by UTurku[16], FBK irst[17] and Kim[3] are all feature-based methods. SCNN[18], CNN[14], SVM-LSTM[19] and MCCNN[15] are all neural network-based methods. It can be seen that the performance of neural network-based methods is generally better than that of feature-based methods, because neural network-based methods can effectively learn useful features automatically. After a comparative analysis of the latest models, the Bert CNN model proposed in this paper has achieved better performance, which is mainly reflected in the following two aspects.

Table 5. An example of a table

Methods	P	R	F1
UTurku[17]	73.20	49.90	59.40
FBK irst[18]	64.60	65.60	65.10
Kim[4]	-	-	67.00
SCNN[19]	72.50	65.10	68.60
SVM-LSTM[20]	75.30	63.70	69.00
CNN[15]	75.70	64.66	69.75
MCCNN[16]	75.99	65.25	70.21
In this paper, methods	74.65	70.74	72.64

First of all, this paper does not use additional NLP tools to obtain features, but automatically obtains features in the process of training. SCNN[18] and SVM-LSTM[19] respectively use Enju[20] and GDep[21] to analyze the syntax of sentences and extract important features. Finally, 68.6% and 69% of F values are obtained. However, these additional tools may not be completely accurate, and may also lead to error propagation, thus hindering the performance of the model.

Secondly, in the existing methods, only word embedding and position embedding are used as features[14, 15], and the semantic information contained in the features only includes word embedding. However, a word may have multiple meanings, so using only words to embed information does not guarantee the correct expression of semantics. In this paper, Bert is used to generate dynamic word vectors to obtain complete word representation.

4. **EXPERIMENT AND ANALYSIS.** In this paper, we propose the Bert-CNN model to extract drug name pairs from biomedical literature. Traditional relationship extraction methods need to extract a large number of features manually, which not only need the participation of professionals, but also rely on natural language processing tools. In this paper, the word vector generated by Bert is used as the input of convolutional neural network, without other features and natural language processing tools. Finally, 72.64% of the F value is obtained on the DDIExtraction2013 data set, which is 2.43% higher than the latest method.

For further study, we will continue to study the improvement of Bert-CNN model, and explore more efficient relationship classifiers to make the model more accurate. We could improve the effect of four kinds of relationship classification by integrating text information or improving the model, so as to obtain better performance.

REFERENCES

[1] Segura-Bedmar I, Martínez P, Pablo-Sánchez C. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinform*, 12(Suppl. 2), S1, 2011.

[2] Wang J, Liu M J, Lin H F. Protein-protein Interaction Extraction Method Based on Multiple Features and Multiple Classifiers Fusion. *Computer Engineering*, vol. 41, no. 11, pp. 207-212, 2015.

[3] Kim S, Liu H, Yeganova L, et al. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. *Journal of Biomedical In-formatics*, 55:23-30, 2015.

[4] Hubel D H , Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, vol. 160, no. 1, pp. 106-154, 1962.

[5] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. *Eprint Arxiv*, 2014.

[6] Mikolov T, Sutskever I, Chen Kai et al. Distributed representations of words and phrases and their com-positionality. *Proceeding of the 26th Advances in Neural Information Processing Systems*, [2017-07-04]. arXiv:1310.4546, 2013.

[7] Pennington J, Socher R, Manning C. Glove:Global vectors for word representation. *Proceeding of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 1532-1543[2017-07-04], 2014.

[8] Peters M E, Neumann M. Deep Contextualized Word Representations. *arXiv*: 1802, 05365v2, 2018.

[9] Radford A, Narasimhan K. Improving Language Un-derstanding by Generative Pre-Training[J/OL]. `https://s3-us-west2.amazonaws.com/openai-assets/research-covers/languagunsupervised/langua\_understanding\_paper.pdf`

[10] Devlin J , Chang M W , Lee K , et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Un-derstanding*, 2018.

[11] S. Marra, M. A. Iachino, and F. C. Morabito, Tanh-like Activation Function Implementation for High-performance Digital Neural Systems, in *Research in Microelectronics & Electronics*, PhD, 2006.

[12] H. K. Kwan, Simple sigmoid-like activation function suitable for digital hardware implementation, *Elec-tronics Letters*, vol. 28, pp. 1379-1380, 1992.

[13] Y. Li, Y. Yuan. *Convergence Analysis of Two-layer Neural Networks with ReLU Activation*, 2017.

[14] Shengyu L, Buzhou T, Qingcai C, et al. Drug-Drug Interaction Extraction via Convolutional Neural Net-works. *Computational & Mathematical Methods in Medicine*, 1-8, 2016.

[15] Chanqin Q , Lei H , Xiao S , et al. Multichannel Con-volutional Neural Network for Biological Relation Extraction. *Biomed Research International*, 1-10, 2016.

[16] Bjorne J, Kaewphan S, Salakoski T. UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, SemEval, NAACL-HLT 2013, Atlanta, GA, USA, 14-15 June 2013.

[17] Faisal M, Chowdhury M, Lavelli A, et al. *FBK-irst: A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information*, 2013.

[18] Zhao Zhehuan, Yang Zhihao, Luo Ling, et al. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network[J]. *Bioinformatics*, (22):22, 2016.

[19] Degen Huang, Zhenchao Jiang, Li Zou, et al. Drug drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. *Information Sciences*, 415, 2017.

[20] Miyao Y, Tsujii J. Feature forest models for probabilistic HPSG parsing. *Comput Linguist*, 34: 35-80, 2008.
[21] Sagae K. *Dependency parsing and domain adaptation with LR models and parser ensembles*, 2007.