I. J. Mathematical Sciences and Computing, **2022, 1, 28-36**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2022.01.03

A Multi-channel Character Relationship Classification Model Based on Attention Mechanism

Yuhao Zhao, Hang Li, Shoulin Yin
Software College, Shenyang Normal University, Shenyang 110034, China
Email: yslinhit@163.com

Received: 18 May 2021; Revised: 16 June 2021; Accepted: 15 July 2021; Published: 08 February 2022

Abstract: Relation classification is an important semantic processing task in the field of natural language processing. The deep learning technology, which combines Convolutional Neural Network and Recurrent Neural Network with attention mechanism, has always been the mainstream and state-of-art method. The LSTM model based on recurrent neural network dynamically controls the weight by gating, which can better extract the context state information in time series and effectively solve the long-standing problem of recurrent neural network. The pre-trained model BERT has also achieved excellent results in many natural language processing tasks. This paper proposes a multi-channel character relationship classification model of BERT and LSTM based on attention mechanism. Through the attention mechanism, the semantic information of the two models is fused to get the final classification result. Using this model to process the text, we can extract and classify the relationship between the characters, and finally get the relationship between the characters included in this paper. Experimental results show that the proposed method performs better than the previous deep learning model on the SemEval-2010 task 8 dataset and the COAE-2016-Task3 dataset.

Index Terms: Relation classification; attention mechanism; BERT; LSTM

1. Introduction

With the increasing abundance of network information resources, information data presents the characteristics of huge scale, diverse modes and high-speed growth. How to extract useful information from massive data quickly and effectively has become a key competitiveness. Google launched knowledge map in 2012 and applied it to search engine to enhance the accuracy of search results, which marks the successful application of large-scale knowledge map in Internet semantic search[1]. Relation classification is to recognize the semantic relationship between two entities in a text[2], which plays an important role in many natural language processing tasks.

There are two kinds of traditional relation classification methods, rule-based method and feature vector based method. The rule-based method needs the intervention of domain experts and needs to build a large number of matching rules manually, which has poor scalability. The feature-based method needs to construct a large number of features manually, which is time consuming and laborious, and the features extracted manually stay at the lexical and syntactic level, so the model can not capture the semantic features of the text well. Stephen[3] et al. proposed a relevance feedback system based on regression prediction model and TF-IDF algorithm, which is a good representative in this aspect.

In recent years, with the development of machine learning and deep learning, Ashwani Kharola[4] proposed a artificial neural network for predicting LVDT output characteristics. In addition, many algorithms and neural networks are first used in computer vision tasks, and have achieved better performance in various industries. Shoulin Yin[5] et al. proposed a active contour model based on density-oriented BIRCH clustering method, which can segment multimedia medical images well. Xiaowei Wang[6] et al. proposed a network intrusion detection method based on deep multi-scale convolutional neural network. Then, a variety of neural network model has been applied to various relation classification tasks. The LSTM can model the hierarchical structure of sentences and solve the problem of long text distance dependence. As a new model, BERT refreshes the best results of many NLP tasks, and then it is applied to text classification and relation classification. However, due to the lack of specific preprocessing and optimization, the effect of character relationship classification is not good. On the basis of these studies, this paper proposes a multi-channel character relationship classification model based on attention mechanism, which combines the semantic information of the two models.

2. Related Work

At present, the existing relation classification methods include: rule based method, feature vector based method, kernel function based method and deep learning model based method.

The rule-based approach relies on domain experts, a large number of pattern matching rules are constructed to classify relation, the task of relation classification is suitable for specific domain. Aone[7] and others developed REES system, which can can identify more than 100 relations. Humphrey[8] et al. complex syntactic rules are constructed to recognize the relation between entities. Through these complex syntactic rules, it can finally get better results. These methods rely on experts in specific fields and can achieve better classification results. However, they are time-consuming, laborious and poor portability.

The feature-based method proposes a series of schemes using various text features, then transforms features into vectors. Using machine learning algorithm to build the model. The feature vector is used as the input of the model to classify the relation between entity pairs. Kambhatla[9] et al., by combining lexical, syntactic and semantic features, the maximum entropy model is used as a classifier, and the feature is transformed into a vector to ensure the final ideal effect. The F value of the final classification is 52.8% at ACE-RDC2003. It is a typical representative of feature-based methods. However, these schemes can not reasonably utilize the structural information (such as word sequence, parsing tree, etc.) in the text.

The method based on kernel function can use a large number of text features without explicitly specifying how to extract text features. The corresponding kernel methods include: Qian L[10] et al. proposed a tree kernel-based semantic relation extraction based on exploiting constituent dependencies. Mooney R J[11] et al. proposed a shortest path dependency kernel for relation extraction, which makes good use of the idea of shortest path dependence. Mintz M[12] et al. proposed a method of unmarked data relation extraction is used for remote monitoring, which is a typical application of using kernel function to extract relational features. However, the biggest problem of supervised learning is that it needs a lot of labeled data. In 2009, Mintz[13] et al. proposed a Distance Supervision (DS) method to generate labeled data, such as extracting two entities from labeled data, searching in unlabeled data, and identifying sentences containing two entities as positive samples, or negative samples, but using DS method to mark the wrong data will bring noise to the training samples. In view of this shortcoming of DS, Riedel and Hoffmann put forward the hypothesis relaxation version of DS. Riedel S[14] et al. proposed a method of modeling relations and mentions without labeled text. Meanwhile, Hoffman R[15] et al. proposed a method of information extraction based on knowledge weak supervision in overlapping relation areas. Takamatsu[16] and others found the shortcomings of DS method with relaxed hypothesis, and then proposed a novel generative model to model the marking process, which successfully reduced the marking errors.

In recent years, thanks to the development of deep learning, CNN is the first deep learning model applied to relational classification tasks, Li[17] et al. combined with dependency tree and hierarchical convolution method, an improved CNN model was proposed and applied to the relation classification task. This method can better apply the convolution idea to relation classification. Santos[18] et al. made improvement on the basis of CNN model, and introduced new loss function to distinguish some categories which are easy to be divided into errors. In the work of Socher[19] et al., they did not use CNN as the basic structure, instead, started to try to use RNN for relationship classification. The possibility of applying RNN in relationship classification was verified. Because CNN can extract the word level features and RNN can extract the sentence level features, Guo[20] et al. proposed the Att-RCNN (attention based combination of CNN and RNN) model, which is combined with RNN and attention mechanism. The application of attention mechanism in relationship classification was proposed. Wang[21] et al. proposed Att-Pooling-CNN model by combining two attention layers with CNN to better identify patterns in heterogeneous contexts, and finally, the classification effect of entity relations reaches a new high level.

3. Methodology

In this paper, a multi-channel model based on attention mechanism is proposed. LSTM and Bert are used as two channels respectively to output their own features, and the features are fused through attention mechanism. The overall structure is shown in Fig.1, which is mainly composed of the following three parts:

- LSTM model channel: Input the text information into LSTM model, mainly deal with the time series features of the input, and output the feature results.
- BERT model channel: Input the text information into pre-trained BERT model, learn the semantic information and understand the essential meaning of the text.
- The feature fusion method based on attention mechanism: Attention mechanism is used to fuse the feature results generated by LSTM channel and BERT channel.

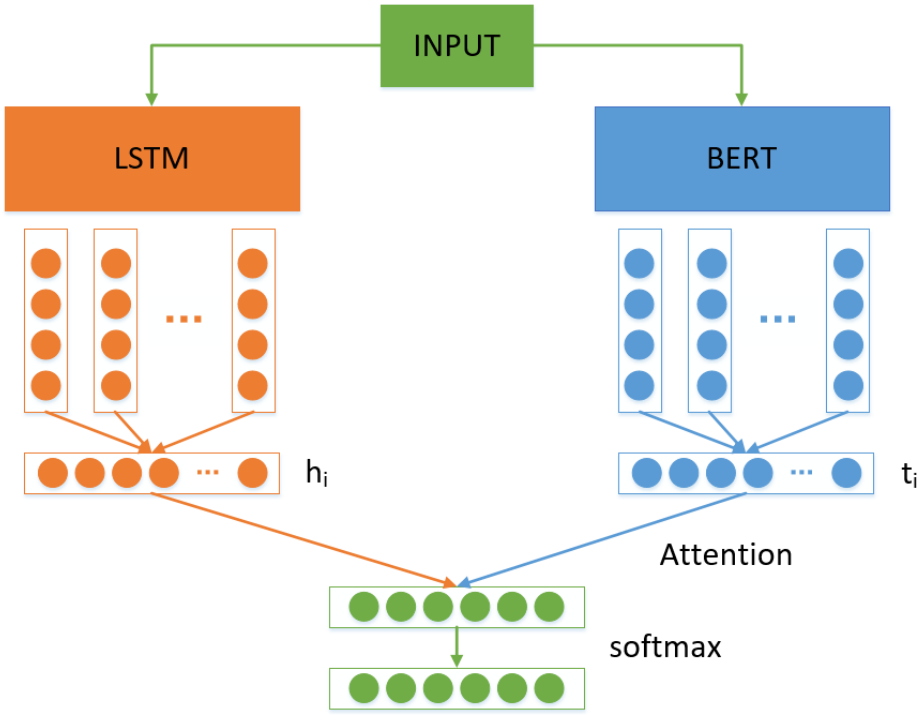

Fig.1. The overall model architecture of LSTM-BERT_Att. The input text passes through two channels: LSTM and Bert. After processing, the attention mechanism is used to fuse the output. Finally, the final result is obtained through softmax function.

3.1 LSTM

Long Short-Term Memory (LSTM) is a special RNN, which is mainly used to solve the problem of gradient disappearance and gradient explosion in the process of long sequence training. LSTM can perform better in longer sequences than RNN. The model architecture of LSTM is shown in Fig.2.

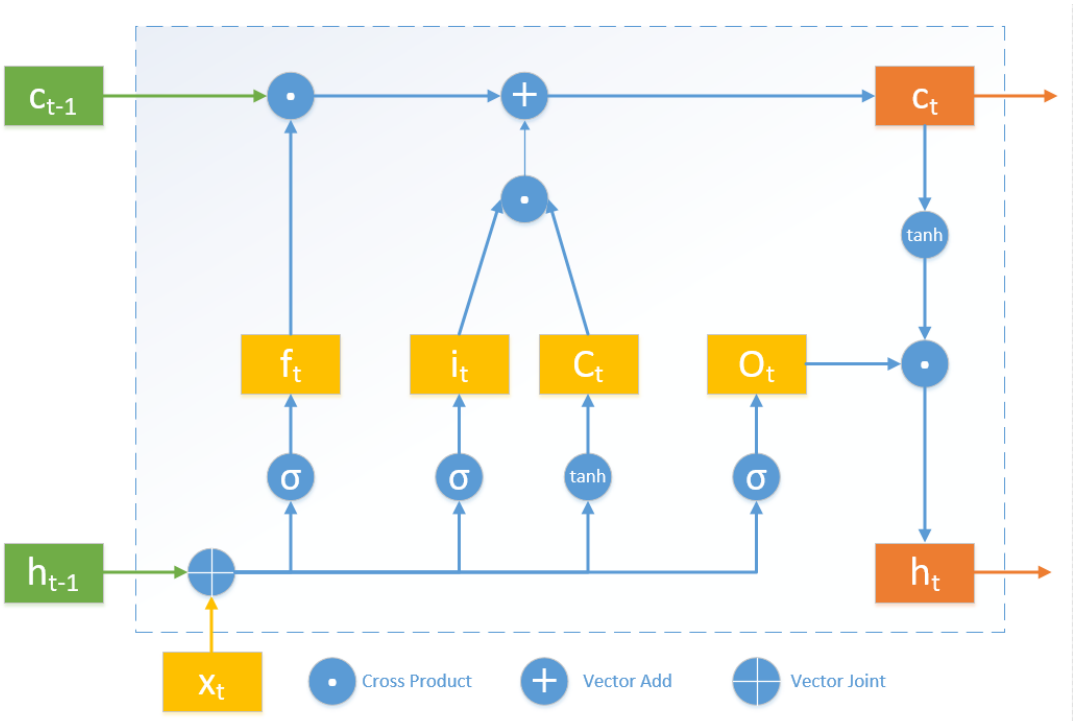

Fig.2. The overall model architecture of LSTM. A gate mechanism is introduced to control the circulation and loss of features. Update the status according to the existing input and the output of the previous cell, and then output the predicted value according to the existing status.

3.2 BERT

Bert is a recently proposed language pre-trained model, which uses the bidirectional transformer model structure[22] to pre-trained large unmarked corpus, and then shows the model performance on some NLP tasks (such as word segmentation, named entity recognition, emotional analysis and problem solving) by fine tuning downstream tasks. The model architecture of Bert is shown in Fig.3.

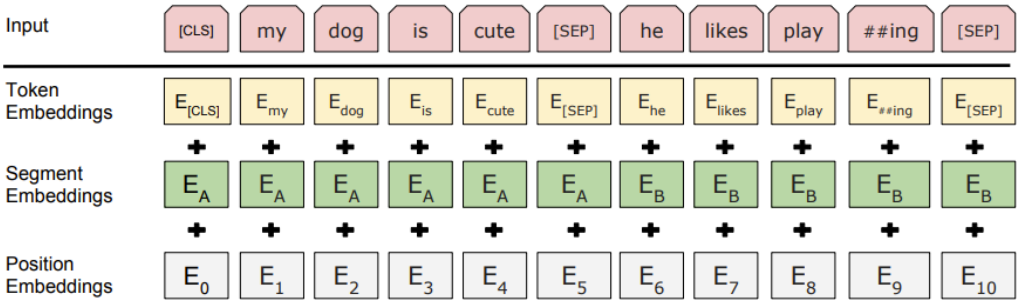

Fig.3. The overall architecture of BERT. Where: Token Embedding is the embedding of words, which is obtained through training and learning; Segment embedding is used to distinguish which sentence each word belongs to, which is also obtained through training and learning; Position embedding is used to encode the position where the word appears.

3.3 Attention Mechanism

In order to fuse the time series features extracted by LSTM and the semantic features learned by Bert, and further understand the implicit semantic expression of text, this paper proposes a semantic fusion method based on attention mechanism. This method combines the feature h_i obtained from LSTM model channel with the semantic t_i obtained from Bert model channel with the method of attention mechanism to obtain the final feature of the text.

Firstly, cosine similarity algorithm is used to measure the correlation between h_i and t_i, and the feature with high correlation will get higher weight:

$$f(h_i^{'}, t_i^{'}) = \cos(h_i, t_i) \tag{1}$$

Then, the correlation is normalized by softmax function to get the attention weight u_i:

$$u_i = \frac{\exp(f(h_i^{'}, t_i^{'}))}{\sum_{i=1}^{n} \exp(f(h_i^{'}, t_i^{'}))} \tag{2}$$

Finally, the fusion feature vector U_i and feature matrix U with rich semantics are obtained by weighted operation of u_i and t_i:

$$U_i = \sum_{i=1}^{n} u_i \bullet t_i^{'} \tag{3}$$

$$U = [U_1, U_1, U_1, \cdots, U_H] \tag{4}$$

4. Experiments

4.1 Experimental Environment

The experimental environment and its configuration are shown in Table 1.

Table 1. Experimental environment configuration

Experimental Environment	Environment Configuration
OS	Ubuntu 16.04
CPU	Intel Core i7-9700K
GPU	NVIDIA GeForce GTX 2070
Memory	32GB
Programming Language	Python 3.6
Deep Learning Framework	Keras 2.2.4

4.2 Experimental Data

Semeval-2010-task8 dataset and coae-2016-task3 dataset were used in this experiment. The two datasets have been widely used in relation classification task, and the classification effect of the model can be verified by comparing with other methods. Semeval-2010-task8 dataset is the main dataset, the "other" category indicates that there is no relation between entity pairs, and various relations are shown in Fig.4. The data set contains 10717 texts, including 8000 texts as training set and 2717 texts as test set.

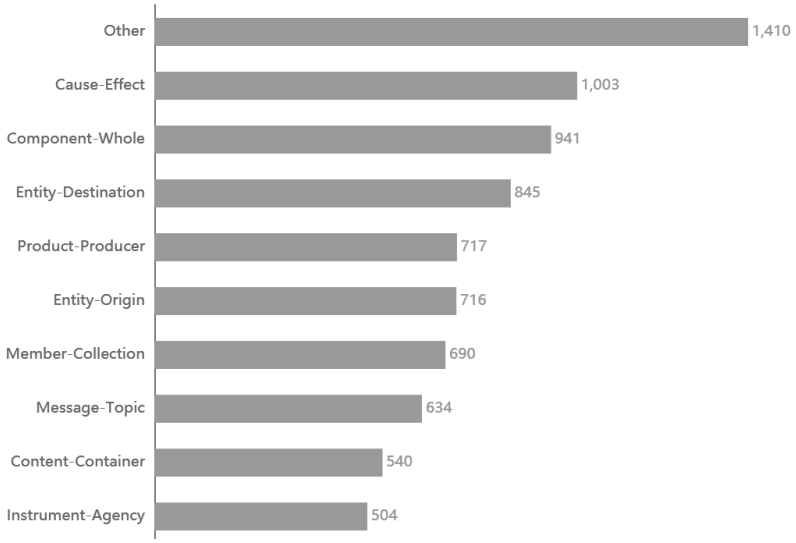

Fig.4. SemEval-2010-Task8 relation type and distribution

4.3 Evaluation

For each relation category, the P (Precision), R (Recall) and F1 (F1_score) value are generally used to measure the classification effect of the model. From the distribution of relation categories in Fig.4, we can see that the data distribution is roughly balanced, so this paper uses macro-F1 to measure the performance of the model. Firstly, the values of P, R and F1 of each class sample are calculated respectively. In equations (5) - (7), i is the class i sample, TP is the number of positive samples correctly predicted, TN is the number of negative samples correctly predicted, FP is the number of positive samples wrongly predicted, and FN is the number of negative samples wrongly predicted.

$$P_{\mathrm{i}} = \frac{TP_i}{TP_{\mathrm{i}} + FP_i} \times 100\% \tag{5}$$

$$R_{\mathrm{i}} = \frac{TP_i}{TP_{\mathrm{i}} + FN_i} \times 100\% \tag{6}$$

$$F1_{\mathrm{i}} = \frac{2 \times P_i \times R_i}{P_{\mathrm{i}} + R_i} \times 100\% \tag{7}$$

The macroscopic precision rate P_{ma} is the mean of the precision rate of all categories, and macroscopic recall rate R_{ma} is also the mean of recall rate of all categories. Macroscopic F1 value $F1_{ma}$ is also the mean value of F1 value of all categories.

$$P_{ma} = \frac{P_1 + P_2 + \cdots + P_N}{N} \tag{8}$$

$$R_{ma} = \frac{R_1 + R_2 + \cdots + R_N}{N} \tag{9}$$

$$F1_{ma} = \frac{F1_1 + F1_2 + \cdots + F1_N}{N} \tag{10}$$

4.4 Parameter

In the training process, the training data is divided into training set and verification machine according to the ratio of 9:1, and the mean value of the verification results is taken as the evaluation of the current model. The experimental results show that when the model is most effective, the main parameters are set as follows. In the process of data preprocessing, the maximum sentence length of the dataset is 220. In the model training stage, the batch of Chinese and English datasets is 16. In order to prevent over fitting, the experiment sets the L2 regularization coefficient λ to 10^{-5}, and finally use Adam algorithm to optimize the update iteration parameters.

4.5 Experiment Results and Analysis

Experiment 1. Comparison of relation classification methods on SemEval-2010-Task8

In this group of experiments, the classification effect of the proposed model is compared with other methods under different feature support. These methods include convolutional neural network correlation methods CNN and CR-CNN, CNN or RNN combined with attention mechanism methods include Att-CNN, BGRU-Att, Att-BLSTM, Att-RCNN, Att-Pooling-CNN and BLSTM-Entity_Att. WE (Word Embedding), POS (Part-of-Speech), NER (Name Entity Recognition), PF (Position Feature) and PI (Position Indicators) are the feature sets used in the above methods.

Table 2 shows that CNN[23] uses the original sentence sequence as input, and uses the location feature to highlight the location information of entity pairs. Obviously, PF is very important for entity relationship classification task, because $F1_{ma}$ increased from 68.4% to 81.6%. CR-CNN[18] considered more the influence of "Other" relation category, and improved the loss function to reach the $F1_{ma}$ value of 83.5%. Attention mechanism has achieved effective results in different fields such as image, text and voice, which makes Att-BLSTM[24], Att-CNN[25] and BGRU-Att[26] achieve better results in relation classification task. For Att-RCNN and Att-Pooling-CNN, they are the best representatives of CNN and RNN methods in relation classification task, with scores of 86.3% and 87.9% of $F1_{ma}$. The model in this paper is a multi-channel model based on attention, using LSTM and Bert at the same time. The experimental data show that 88.5% of the macroscopic $F1_{ma}$ value is achieved on SemeVal-2010-Task8.

Table 2. Comparison of relation classification methods on SemEval-2010-Task8

Model	Feature Sets	$F1_{ma}$ (%)
CNN	WE	68.4
	+PF, WordNet	81.6
CR-CNN	WE + PF	83.5
Att-BLSTM	WE, PI	84.2
BLSTM-Entity_Att	WE, Latent Entity Typing	85.0
Att-CNN	WE	85.8
BGRU-Att	WE, PF, POS, NER, etc	85.9
Att-RCNN	WE	86.3
Att-Pooling-CNN	WE	87.9
LSTM-BERT_Att	WE, Positional Encoding	88.5

Experiment 2. Comparison of relation classification methods on COAE-2016-Task3

In Table 3, ET represents the entity type, EO represents the entity order, and * indicates that the paper data is not listed. The experimental data show that P_{ma} and R_{ma} in CNN model are relatively balanced, but the effect is not good due to the small number of training samples in COAE-2016-Task 3. When the training samples are large, CNN

still has a lot of room to improve. Compared with CNN model, PCNN_ATT improves the $F1_{ma}$ by 11.33% because it introduces the word level attention mechanism and adopts the strategy of segmented maximum pooling operation. At the same time, it also proves that ET is helpful to distinguish some relation categories.

Table 3. Comparison of relation classification methods on COAE-2016-Task3

Model	Feature	P_{ma} (%)	R_{ma} (%)	$F1_{ma}$ (%)
CNN	WE, PF	59.99	55.87	57.03
PCNN	WE, PF	68.45	62.32	64.89
PCNN_ATT	WE, PF	75.38	66.89	68.36
PCNN_ATT	WE, PF, ET	77.64	76.23	76.58
SelfAtt-BLSTM	WE, POS, PF	*	*	84.20
LSTM-BERT_Att	WE, PI	91.10	90.38	92.03

Compared with PCNN_ATT and SelfAtt-BLSTM model, the model proposed in this paper can greatly improve the classification effect on Chinese entity relation classification task COAE-2016-Task3. The above reasons make the model in this paper achieve the best effective on COAE-2016-Task3, and finally achieve 92.03% of the $F1_{ma}$ value.

5. Conclusion

In this paper, the LSTM model is used to obtain the sequence feature information of the text, and the semantic feature information of the input text is obtained through the pre-trained model Bert. Then, the time series feature and semantic feature are fused through the attention mechanism, and finally the results are sent to the softmax function to complete the classification task. The model does not need any background knowledge and syntactic features as auxiliary information, and only uses the original text as input to make more effective use of the information of the input text. By inputting the original text in the dataset into this model, the relationship between characters contained in the text can be extracted and classified, and the accuracy of the final result has a good performance. Experimental results on SemEval-2010-Task8 and COAE-2016-Task3 show that the proposed model has better performance for relation classification tasks.

However, it is found that the method proposed in this paper is not effective for cross domain text classification, and the next research work will focus on this to further improve the classification effect of the model for this kind of problems.

References

[1] Ahmet Uyar, Farouk Musa Aliyu. "Evaluating search features of Google Knowledge Graph and Bing Satori" Online Information Review, 2015, 39(2).

[2] Wang L, Cao Z, De Melo G, et al. "Relation classification via multi-level attention CNNS" Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1298-1307.

[3] Stephen Akuma,Rahat Iqbal. Development of Relevance Feedback System using Regression Predictive Model and TF-IDF Algorithm[J]. International Journal of Education and Management Engineering(IJEME),2018,8(4).

[4] Ashwani Kharola. Artificial Neural Networks based Approach for Predicting LVDT Output Characteristics[J]. International Journal of Engineering and Manufacturing(IJEM),2018,8(4).

[5] Shoulin Yin, Hang Li*, Desheng Liu and Shahid Karim. "Active Contour Model Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation" Multimedia Tools and Applications. Vol. 79, pp. 31049-31068, 2020.

[6] Xiaowei Wang, Shoulin Yin, Hang Li. "A Network Intrusion Detection Method Based on Deep Multi-scale Convolutional Neural Network." International Journal of Wireless Information Networks. 27(4), 503-517, 2020.

[7] Aone C, Ramos-Santacruz M. REES: A Large-Scale Relation and Event Extraction System[J]. proceedings of anlpnaacl, 2002.

[8] Humphrey Susanne M,N évéol Aur élie,Gobeil Julien,Ruch Patrick,Darmoni St éfan J,Browne Allen. Comparing a Rule Based vs. Statistical System for Automatic Categorization of MEDLINE Documents According to Biomedical Specialty.[J]. Journal of the American Society for Information Science and Technology : JASIST,2009,60(12).

[9] Kambhatla N. "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations" Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004:22.

[10] Qian L, Zhou G, Kong F, et al. "Exploiting constituent dependencies for tree kernel-based semantic relation extraction" Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for computational Linguistics, 2008:697-704.

[11] Mooney R J, Bunescu R C. "Subsequence kernels for relation extraction" Advances in neural information processing systems. 2005: 171-178.

[12] Bunescu R C, Mooney R J. "A shortest path dependency kernel for relation extraction" Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005: 724-731.
[13] Mintz M, Bills S, Snow R, et al. "Distant supervision for relation extraction without labeled data" Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 1003-1001.
[14] Riedel S, Yao L, McCallum A. "Modeling relations and their mentions without labeled text" Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010: 148-163.
[15] Hoffman R, Zhang C, Ling X, et al. "Knowledge-based weak supervision for information extraction of overlapping relations" Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 541-550.
[16] Takamatsu S, Sato I, Nakagawa H. "Reducing wrong labels in distant supervision for relation extraction" Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 721-729.
[17] Li B, Zhao X, Wang S, et al. "Relation classification using revised convolutional neural networks" 4th International Conference on Systems and Informatics(ICSAI), 2017: 1438-1443.
[18] Santos C N, Xiang B, Zhou B. "Classifying relations by ranking with convolutional neural networks" Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics(ACL), 2015: 626-634.
[19] Socher R, Huval B, Manning C D, et al. "Semantic compositionality through recursive matrix-vector spaces" Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics(EMNLP-CoNLL), 2012: 1201-1211.
[20] Guo X, Zhang H, Yang H, et al. "A single attention -based combination of CNN and RNN for relation classification." IEEE Access, 2019, 7(1): 12467-12475.
[21] Wang L, Cao Z, De Melo G, et al. "Relation classification via multi-level attention cnns" Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(ACL), 2016: 1298-1307.
[22] Vaswani A, Shazeer N, Pamar N, et al. "Attention is all you need" Advances in Neural Information Processing Systems, 2017: 5998-6008.
[23] Zeng D, Liu K, Lai S, et al. "Relation classification via convolutional deep neural network" 25th International Conference on Computational Linguistics(COLING), 2014: 2335-2344.
[24] Zhou P, Shi W, Tian J, et al. "Attention-based bidirectional long short-term memory networks for relation classification" Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(ACL), 2016: 207-212.
[25] Shen Y, Huang X. "Attention-based convolutional neural network for semantic relation extraction" Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers(COLING), 2016: 2526-2536.
[26] Meng B, Xu Bao-min, Zhou E, et al. "Bidirectional gated recurrent unit networks relation classification with multiple attentions and semantic information" The 16th International Symposium on Neural Network, 2019: 124-132.

Authors' Profiles

Yuhao Zhao graduted with a Bachelor of Engineering from Changzhi University in 2015. In his college, after completing the learning task, he interests in exploring his professional knowledge. Now, he is a undergraduate of Master of Engineering. His research interests include nature language processing and image processing.

Hang Li obtained his Ph.D. degree in Information Science and Engineering from Northeastern University. Hang Li is a full professor of the Software college at Shenyang Normal University. He is also a master's supervisor. He has research interests in wireless networks, mobile computing, cloud computing, social networks, network security and quantum cryptography. Prof. Li had published more than 30 international journal and international conference papers on the above research fields.

Shoulin Yin received the B.Eng. and M.Eng. Degree from Shenyang Normal University, Shenyang, Liaoning province, China in 2016 and 2013 respectively. Now, he is a doctor in Harbin Institute of Technology. His research interests include Multimedia Security, Network Security, Filter Algorithm, image processing and Data Mining.

How to cite this paper: Yuhao Zhao, Hang Li, Shoulin Yin," A Multi-channel Character Relationship Classification Model Based on Attention Mechanism ", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.8, No.1, pp. 28-36, 2022. DOI: 10.5815/ijmsc.2022.01.03