Triple Extraction Technique for Power Transformer Fault Information Disposal Based on a Residual Dilate Gated Convolution and Self-Attention Mechanism

HongYing He *, XiHao Yin, DianSheng Luo, RuiYao Xi, Jie Fang, FangYu Fu and GuangWei Luo

College of Electrical and Information Engineering, Hunan University, Changsha, China

This article presents a triple extraction technique for a power transformer fault information process based on a residual dilate gated convolution and self-attention mechanism. An optimized word input sequence is designed to improve the effectiveness of triple extraction. A residual dilate gated convolution is used to capture the middle-long distance information in the literature. A self-attention mechanism is applied to learn the internal information and capture the internal structure of input sequences. An improved binary tagging method with position information is presented to mark the start and the end of an entity, which improves the extraction accuracy. An object entity is obtained by a specific relationship r for a given subject. The nearest start-end pair matching the principle and probability estimation is applied to acquire the optimal solution of the set of triples. Testing results showed that the $F1$ score of the presented method is 91.98%, and the triple extraction accuracy is much better than the methods of BERT and Bi-LSTM-CRF.

Keywords: power transformer, optimized character input sequence, residual dilate gated convolution, self-attention mechanism, binary tag framework, triple

1 INTRODUCTION

At present, the treatment for power transformer faults mainly depends on the subjective decision-making by maintenance personnel who analyze the status and parameter changes of the transformer, find out the causes of faults, and formulate fault processing measures. This method requires maintenance personnel to memorize a large amount of fault processing information in non-structured text such as system stability requirements and key points of fault handling measures. It is significant for the power system to extract knowledge from non-structured fault text by intelligent technology and organize them into structured and visual representation. It will benefit the maintenance personnel to quickly analyze the cause of the faults, comprehensively make fault treatment, and improve the handling ability of emergency of the power transformer.

The concept of knowledge map was mentioned by Google in 2012. A knowledge map is a kind of typical multilateral relationship graph which consists of the entity—relation—entity triplet. The 'entity' is the basic element of the knowledge map, 'relationship' is the relationship between different entities, 'property' is the description of the entity, and 'value' is the value of the entity attributes. The knowledge map has become one of the basic technologies of semantic search (Wickramarachchi

et al., 2021), intelligent question and answer system (Yanchao et al., 2017), knowledge reasoning (Pal., 2020), and other knowledge intelligent services.

Entity recognition and relation extraction algorithms based on deep learning are hot issues of the current research. On entity recognition, Qiu et al. (2019) proposed a model combining the residual dilatation convolutional neural network and CRF, in which RD-CNN is used to capture context features, and CRF is used to capture the correlation of adjacent tags. Kong et al. (2021) combined the attention mechanism with the CNN; different convolutional kernels are integrated to capture context information, and the attention mechanism is applied to enhance the capture ability. Huang et al. (2015) proposed Bi-LSTM and Bi- LSTM-CRF models. In Ma and Hovy. (2016), word-level and character-level embedding vectors are simultaneously integrated as the input of Bi-LSTM-CRF, which achieves good recognition. The aforementioned literature studies are cases on entity recognition; as for relation extraction, Xiaoyu et al. (2019) proposed an ATT-RCNN model to extract text features and accomplish text classification, in which the attention mechanism is used to strengthen features and improve classification performance. Shu et al. (2015) proposed Bi-LSTM for relationship classification which confirmed the effectiveness of Bi-LSTM on relation extraction. The joint extraction method combines entity recognition and relationship extraction to extract triples. Miwa and Bansal. (2016) extracted the relationship between the word sequences and the entities based on the independence structure of the sequence, and Bi-LSTM coding layer information is shared by entity recognition and relationship extraction.

This article presents a triple extraction algorithm for a power transformer fault information process based on a residual dilate gated convolution and self-attention mechanism. First, an optimized character input sequence is designed, and a position vector is added to make the word input sequence have position information. Then, a residual dilate gate convolution network is designed to extract the features of the input sequences, and a self-attention mechanism is applied to capture the internal information and the structure of input sequences. Entities are tagged by an improved binary tag method. Object entities are obtained by a specific relationship object for a given subject, and triples are obtained by the optimized probability estimation. Testing results show that triples are successfully extracted by the presented method.

2 TRIPLE EXTRACTION MODEL BASED ON THE RESIDUAL DILATE GATE CONVOLUTION AND ATTENTION MECHANISM

The architecture of the triple extraction model based on the residual dilate gate convolution and attention mechanism proposed in this study is shown in **Figure 1**, which mainly includes the optimized character input sequence, the residual dilate gate convolution layer (RD-GCNN), self-attention layer, and the entity relation extraction structure by the binary tag method.

2.1 Optimized Character Input Sequence

As the basic unit of input, character can avoid the wrong segmentation of the boundary to the greatest extent, but a simple character vector is difficult to store effective semantic information; character-word mixed coding method is more effective on semantic information. An optimized character input sequence (OCIS) is designed in this article. First, the Jieba (Huiqin and Weiguo., 2018) or HanLP (Wu et al., 2020) segmentation system has the function of adding a custom dictionary, and adding the power field dictionary can solve the problem of excessive segmentation of power field words in the segmentation system. Then, the Word2Vec model (Lu and Jing., 2018; Amin et al., 2020; Dou et al., 2021) was used to train a word vector and establish the fixed word vector query sequence $W = [w]_1^m$. Given a transformer sentence L, the character vector sequence $X = [x]_1^n$ was obtained by an embedding layer, and $x_i \in R^{d_x}$. Jieba or HanLP was used for sentence L segmentation and query corresponding word vector w_i in W for words in L. The word "变压器(transformer)" was taken in the sentence in **Table 6** as an example; the character-word mixing method is as follows,

$$Z = [x]_1^3 + [w_1]^3 \times H. \tag{1}$$

In **Eq. 1**, H is the transformation matrix, $H \in R^{d_x \times d_w}$, $Z \in R^{3 \times d_x}$, $[x]_1^3$ is a character sequence for the transformer with the length of 3, and w_i of $[w_1]^3$ is the vector of the word "变压器(transformer)".

This article introduces the position vector (Ashish et al., 2017) to enrich the input information, shown as follows:

$$\begin{cases} PE_{2i}(pos) = \sin\left(pos / 10000^{2i/d_x}\right), \\ PE_{2i+1}(pos) = \cos\left(pos / 10000^{2i/d_x}\right), \end{cases} \tag{2}$$

where d_x is the character sequence dimension, d_{pos} is the position, and i is the dimension. The OCIS is as follows:

$$X = Z^n + PE^n, \tag{3}$$

where $X \in R^{n \times d_x}$, $Z^n \in R^{n \times d_x}$, and $PE^n \in R^{n \times d_x}$.

2.2 Residual Dilate Gate Convolution Network

In the RD-GCNN, the dilate convolution captures the middle-long distance features in the context. The residual structure enables multi-channel transmission of the information, and the gated linear unit (GLU) reduces gradient dispersion and accelerate convergence. Convolution layers are used to extract the characteristic information of the input data. With the increase in the convolution layer, the context features output by each output element become richer.

2.2.1 Dilated Convolution

Figure 2 shows a one-dimensional dilate convolution with expansion coefficients of 1, 2, and 4. The structure of dilated convolution differs from the standard convolution; here, the dilated convolution captures the information of the central
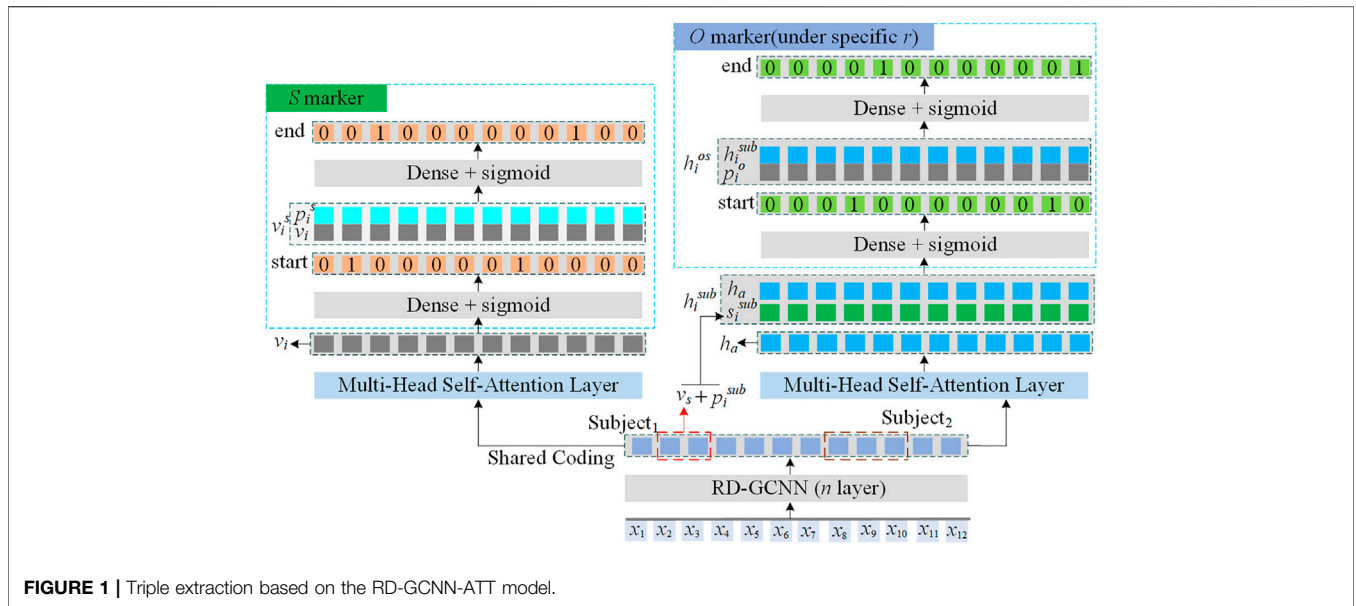
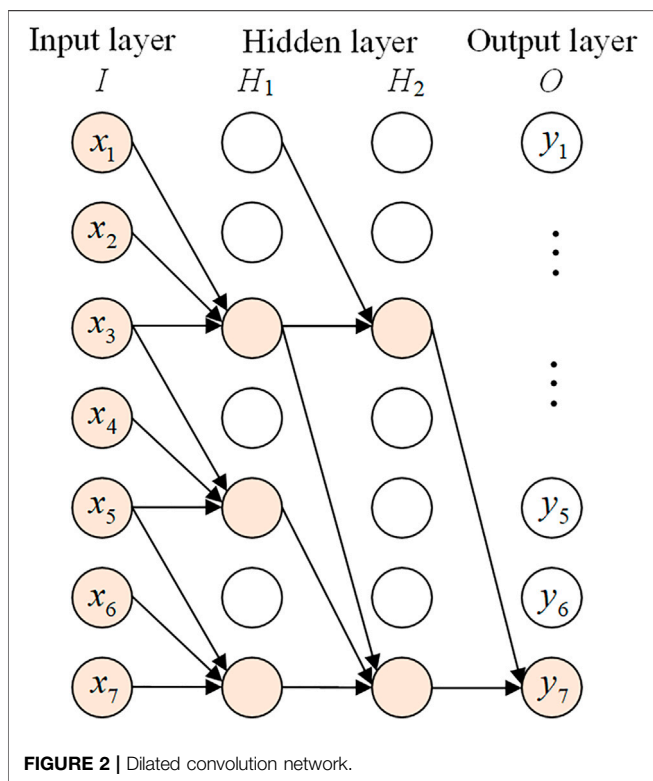**FIGURE 1 |** Triple extraction based on the RD-GCNN-ATT model.

FIGURE 2 | Dilated convolution network.

node and its adjacent nodes at the hidden layer H_1, while the hidden layer H_2 captures the information of nodes by skipping every three inputs consecutively. Thus, features in a longer distance can be captured by changing the expansion coefficients d, and the larger vision of the perceived field is achieved. The local receptive field size is equal to $(k-1)*d+1$, where k is the size of the convolution core. With the increase in the number of dilated convolutional layers, the range of historical

data covered by the convolutional network can be exponentially expanded. For an input sequence X, given the convolution kernel with size k and weight matrix $[w]_0^{2l}$, the dilate convolution equation is as follows:

$$\begin{cases} l = (k-1)/2, \\ s_i = \sum_{j=0}^{2l} w_j \cdot x_{i+j \times d} + b, \end{cases} \quad (4)$$

where k represents the size of the convolution kernel, d is the expansion coefficient, and $w_j \in [w]_0^{2l}$, $x_i \in X$, and b is the bias term. Weight and bias are trained through a large amount of data; during initial training, they are generally randomly generated, and then, errors are constantly searched through the backpropagation algorithm to return the errors step by step to the convolution layer for continuous circulation and correction.

2.2.2 Residual Connection and the Gated Linear Unit
The residual structure can generate a channel for the original information so that the original information can be directly transmitted to the higher layer. GLU is a gating mechanism, and it is best to use GLU as an activation function in natural language processing. Structurally, this part of the convolution without the activation function is not easy to trap into gradient disappearance; the risk of DGCNN gradient disappearing was lower. **Figure 3** shows the dilated convolution network with a residual structure and GLU. The equation is as follows:

$$\delta = \sigma\left(\text{Conv1D}_2(X)\right), \quad (5)$$
$$Y = X \otimes (1 - \delta) + Conv1D_1(X) \otimes \delta, \quad (6)$$

where X is the input, Conv1D_1 and Conv1D_2 are the dilate convolution functions, in which d and k are the same, but the weight is not shared.
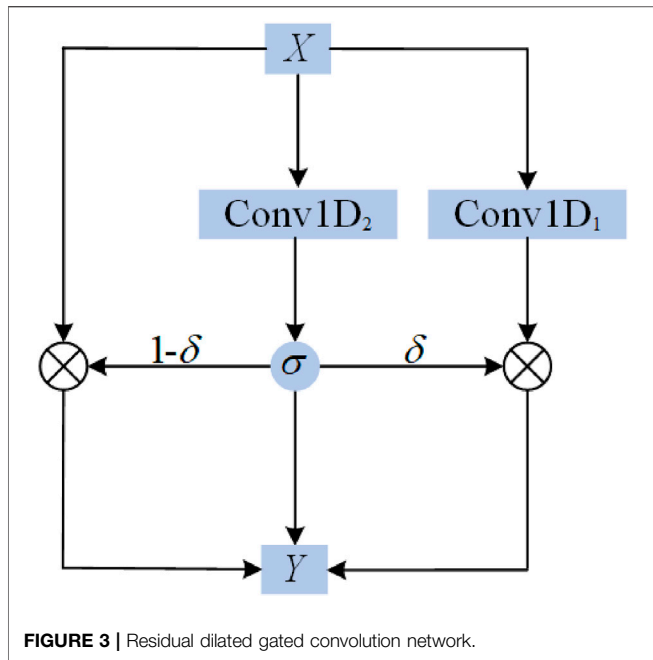
FIGURE 3 | Residual dilated gated convolution network.

2.3 Multi-Head Self-Attention Mechanism

The RD-GCNN is used to capture medium-long term context information, and the attention mechanism (Ashish et al., 2017) is applied to capture the global information, which improves the model performance.

The multi-head self-attention mechanism performs attention calculation for x_i of the input sequence X and all x_i of the input sequence X, with the purpose of learning the dependence relationship of vector x_i inside the input sequence X and capture the global information. The equation is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

$$head_i = \text{Attenion}\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (8)$$

$$\text{Multihead}(Q, K, V) = \text{Concat}\left([head]_1^h\right)W^O, \quad (9)$$

$$Y = \text{Multihead}(X, X, X), \quad (10)$$

where $W_i^{Q,K,V} \in R^{d_x \times d_k}$, $W^O \in R^{d_x \times hd_v}$, factor $\sqrt{d_k}$ plays the adjusting role, Q is a set of queries, K is keys, and V is values, $Q = K = V = X$.

2.4 Entity Relation Extraction Structure of the Binary Tag

The sentences in **Table 6** have the following triples, as shown in **Table 1**. It can be found that the sentences have overlapping relational triples, which directly challenges the traditional sequential labeling scheme. **Figure 1** shows the S marker and O marker (under specific r). First, the subject is tagged from the input sentence; then, for each candidate subject, all possible relations are checked to find possible associate objects in the sentence.

The S marker marks the start-end position of subjects by decoding the encoded vector v_i produced by the RD-GCNN-ATT encoder. The marking process is obtained by **Eqs. 11–17**. Subjects are identified by **Eq. 18**

$$p\left(i^{s-start}\right) = \text{sigmoid}\left(f\left(v_i\right)\right), \quad (11)$$

$$s\_start(x_i) = \begin{cases} 1, p\left(i^{s-start}\right) \geq threshold, \\ 0, p\left(i^{s-start}\right) < threshold, \end{cases} \quad (12)$$

$$p_i^s = \begin{cases} 0, i = i^{s-start} \\ i - i^{s-start}, i > i^{s-start} \\ -1, i < i^{s-start}, \end{cases} \quad (13)$$

$$v_i^s = \text{concat}\left(v_i, p_i^s\right), \quad (14)$$

$$p\left(i^{s-end}\right) = \text{sigmoid}\left(f\left(v_i^s\right)\right), \quad (15)$$

$$s\_end(x_i) = \begin{cases} 1, p\left(i^{s-end}\right) \geq threshold, \\ 0, p\left(i^{s-end}\right) < threshold, \end{cases} \quad (16)$$

$$p(s|m) = \prod_{t \in \{s\_start, s\_end\}} \prod_{i=1}^{L} \left(p\left(i^t\right)\right)^{I\{t(x_i)=1\}} \left(1 - p\left(i^t\right)\right)^{I\{t(x_i)=0\}}, \quad (17)$$

where $f()$ is the dense function, concat () is the splicing function. If $p(\cdot)$ is higher than the *threshold*, the tag of position is set to 1, otherwise 0. p_i^s is the relative position vector of $s\_start(x_i)$, and i is the position of the current character. Adding p_i^s can be a useful indicator to distinguish entity tags from non-entity tags and enhance the extraction of entity and relation. In the process of $s\_start(x_i)$ prediction, the position information before $s\_start(x_i)$ is set to −1, indicating $s\_end(x_i)$ cannot precede $s\_start(x_i)$. Abundant local information can improve the prediction effect. L is the length of the sentence m. If z is true, $I\{z\} = 1$, otherwise 0. $p(s|m)$ represents the maximum joint probability of s_start and s_end for all tokens. In the training process, subjects and objects are real subjects and objects, but in reasoning, predicted subjects are used.

The O marker simultaneously identifies the objects by the involved relations related to the subjects obtained by the S marker. As **Figure 1** shows, it consists of a set of O markers with same structures like S markers. Moreover, it considers subject features v_s and the position information p_i^{sub} of the subject. The O marker tags the start-end of the object like the steps of the tagging of the subject.

As for a tag, the sentence m_i comes from the training set M and a set of triple $T_i = \{(s, r, o)\}$; in the sentence m_i, objects entities are obtained by maximizing joint probability estimation

TABLE 1 | Triplet instance.

Subject	Relation	Object
引线桩头螺丝	发生	松动
Lead pile head screw	Occur	Loose
引线桩头螺丝	发生	发热
Lead pile head screw	Occur	Heat
发热	原因是	引线桩头螺丝松动
Heat	Reason	Lead pile head screw is loose
变压器	包括	引线桩头螺丝
Transformer	Include	Lead pile head screw

TABLE 2 | Character vector.

Parameters	Values
Max sequence length	300
Embedding size	128

TABLE 3 | Word vector.

Parameters	Values
Skip gram	True
Embedding size	200
Window	5
Negative sampling	True
Number of negative sampling	5

TABLE 4 | RD-GCNN.

Parameters	Values
d_x	$d_x = 128$
d_d	$d_d = 128$
d_l	$d_l = 300$
k	$k = 3$
layer	$n = 12$
d	$d = [1,2,5]^3+[1]^3$

presented in **Eq. 18** under specific relationship r, and the principle of the closest matching principle is used.

$$\prod_{i=1}^{|M|}\left[\prod_{(s,r,o)\in T_i} p\left((s,r,o)|m_i\right)\right] = \prod_{i=1}^{|M|}\left[\prod_{s\in T_i} p(s|m_i)\prod_{r\in T_i|s} p_r(o|s,m_i)\right.$$
$$\left. \times \prod_{r\in R\backslash T_i|s} p_r(o_\varnothing|s,m_i)\right],$$
(18)

where $s \in T_i$ is the subject s in the triple T_i, $T_i|s$ means triple T_i is decided by subject s, $(r,o) \in T_i|s$ is a pair of (r,o) in the triple $T_i | s$, $R\backslash T_i|s$ is all relationship r not decided by the subject s, and $o_\varnothing$ is an empty object.

3 EXPERIMENTAL ENVIRONMENT SETTING

Parameters in this article are set according to the experimental experience and process. **Table 2** and **Table 3** show the parameter settings of the optimizable character input sequence.

In this article, the expansion rate was set as a zigzag structure [1,2, and 5] to prevent the grid effect and repeated three times. Finally, [1,1, and1] is used as fine-tuned, as shown in **Table 4**. Parameters of multi-head self-attention are shown in **Table 5**.

TABLE 5 | Multi-head self-attention.

Parameters	Values
d_x	128
h	8
d_k	16
d_v	16
Layer	$n = 1$

4 EXPERIMENT

4.1 Experimental Data

So far, there is no constructed corpus that can be used for the triplet extraction of the power transformer. In order to verify the validity of the model constructed in this article, the data set for model training was obtained by manual annotation from transformer basic information, transformer online monitoring data, transformer fault field maintenance text records, fault case summary, and expert experience text. The model was randomly divided by 8:2 ratio for training and testing.

4.2 Experimental Duties

The task of this article was to extract the transformer equipment type, equipment parts, equipment location, defect phenomenon, defect causes, and maintenance measures. Then, the relationships should be identified between them, such as belongs to, includes, occurs, cause, and measure. As it is shown in **Table 6**, the result of the participle in sentence L "变压器引线桩头螺丝松动导致连接处发热 (loose screw on pile head of transformer lead leads to heat in connection)," the subject entity marker tags the entity "变压器 (transformer)," "引线桩头螺丝 (lead pile head screw)," and "发热 (heat)," the object entity marker marks the "引线桩头螺丝 (lead pile head screw)" specific relationship "发生 (occur)" with the object entity "松动 (loose)," and "发热 (heat)".

4.3 Experimental Results and Analysis

In this study, accuracy p, recalls R, and $F1$ scores are used as evaluation indexes for the effect of the triplet extraction model.

1) In order to verify the effectiveness of the optimizable character input sequence proposed in this article. The optimized character input sequence (OCIS), character and word-mixed sequence (CWMS), character sequence (CS), and word sequence (WS) were introduced into RD-GCNN-ATT for training, and the results are shown in **Table 7**.

It is seen from **Table 7** that the CS input can decrease the influence to the model caused by the wrong word segmentation results. Compared with the CS input, the combined input of CWMS improves the effect by 1–2%, for it has more semantic. The OCIS has the best effect for the addition of location information, and WS input is the worst one in the selected four methods.

TABLE 6 | Triple extraction of the power transformer.

Character sequence	Loose screw on the pile head of the transformer leads to heat in connection																	
Character Sequence	变	压	器	引	线	桩	头	螺	丝	松	动	导	致	连	接	处	发	热
Jieba/HanLP	变压器			引线桩头螺丝						松动		导致		连接处			发热	
English	transformer			lead pile head screw						loose		lead to		in connection			heat	
Subject-start Tag	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Subject-end Tag	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Subject entity	Entity 1			Entity 2						None							Entity 3	
	A random s is selected, and for each relationship r, the corresponding o and hypothesis extraction to Entity 2 are extracted																	
Object-start Tag	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
Object-End Tag	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Object Entity	None									Entity 1				None			Entity 2	
Triples	(引线桩头螺丝, 发生, 松动), (引线桩头螺丝, 发生, 发热)																	
English	(lead pile head screw, occur, and loose), (lead pile head screw, occur, and heat)																	

TABLE 7 | Input sequence experimental results.

Input	p	R	F1
OCIS	0.9244	0.9152	0.9198
CWMS	0.9203	0.9083	0.9143
CS	0.9068	0.8855	0.8960
WS	0.8677	0.8506	0.8591

TABLE 8 | Results of the model experiment.

Model	p	R	F1
RD-GCNN-ATT	0.9244	0.9152	0.9198
RD-GCNN	0.8901	0.8965	0.8933

TABLE 9 | Results of the model experiment.

Model	p	R	F1
IBM + RD-GCNN-ATT	0.9244	0.9152	0.9198
IBM + BERT	0.9086	0.9064	0.9075
UBM + RD-GCNN-ATT	0.9063	0.9130	0.9096
UBM + BERT	0.8920	0.8905	0.8913

TABLE 10 | Results of the model experiment.

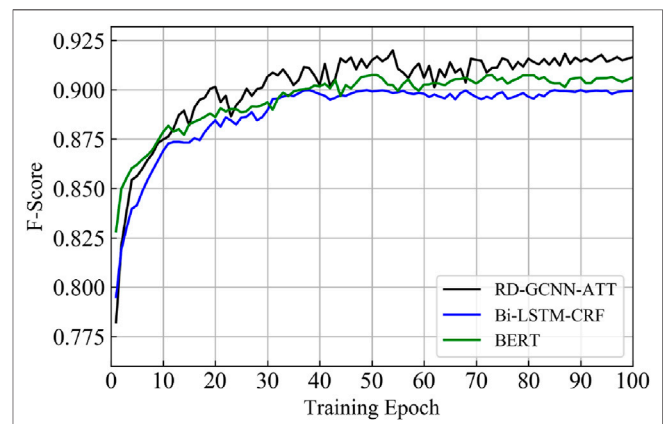| Model	p	R	F1
RD-GCNN-ATT	0.9244	0.9152	0.9198
Bi-LSTM-CRF	0.8965	0.9031	0.8998
BERT	0.9086	0.9064	0.9075

FIGURE 4 | Model training curve.

2) In order to verify the effectiveness of the attention mechanism in RD-GCNN-ATT, the attention mechanism was deleted from the model structure, and the results are shown in **Table 8**.

Dilated convolution can only capture the medium-long term context information, but the attention introduced has a stronger ability to capture the global context information. **Table 8** shows that the attention improves the model performance of sequence labeling.

3) In order to verify the effectiveness of the improved binary marking method, the improved binary marking method (IBM) and unimproved binary marking method (UBI) are used, respectively, to test the extraction effect of RD-GCNN-ATT and BERT models, and the results are shown in **Table 9**.

Table 9 shows that the indexes of the IBM are higher than those of the UBI in RD-GCNN-ATT and BERT models. In the IBM, the relative position vector of the start tag is added to the end tagging of the subject entity tag and the object entity tag, which considers the start tag information and improves the tagging performance. When the subject entity encoding passed into the object entity marker, it considers the position information of the subject, prevents the loss of the subject entity features and improves the performance of tag.

4) In order to verify the effectiveness of the model RD-GCNN-ATT proposed in this article, the testing results of BI-LSTM-CRF and BERT are shown in **Table 10** and **Figure 4**.

As it is seen from **Table 10** and **Figure 4**, the RD-GCNN-ATT model achieves the best performance. The poor performance of the Bi-LSTM-CRF model lies in the fact that it uses CRF sequence annotation and cannot effectively solve the situation of overlapping triples. The performance of the BERT model is the mediate level.

5 CONCLUSION

This article presents a triple extraction algorithm for power transformer fault disposal based on the residual dilate gated convolution and self-attentional mechanism. An optimized character input sequence is designed, and a position vector is added to make the word input sequence have position information. A residual dilate gate convolution network is developed to extract the features of input sequences in a larger range, and a self-attention mechanism is presented to learn the internal dependencies of input sequences and capture the internal structure of input sequences. An improved binary tag method is presented, and an entity marker is trained to tag the start and the end of the entities with the position information. Object entities are extracted by the specific relationship for a given subject, and triples are obtained by the optimized probability estimation. Compared with the testing results of the Bi-LSTM-CRF and BERT models, the presented method has the best accuracy on triple extraction of the transformer fault information.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HH: Conceptualization, methodology, supervision, validation, and writing review and revision. XY: Gathering data, debugging programs, and writing original draft and editing. All authors contributed to the manuscript and approved the submitted version.

FUNDING

REFERENCES

Amin, S., Uddin, M. I., Zeb, M. A., Alarood, A. A., Mahmoud, M., and Alkinani, M. H. (2020). Detecting Dengue/Flu Infections Based on Tweets Using LSTM and Word Embedding. *IEEE Access* 8, 189054–189068. doi:10.1109/ACCESS.2020.3031174

Ashish, V., Noam, S., and Niki, P. (2017). "Attention Is All You Need," in 31st Conference on Neural Information Processing Systems (NIPS 2017) (Long Beach, CA, USA: arXiv:1706.03762).

Dou, Z., Cheng, Z., and Huang, D. (2021). Research on Migrant Works' Concern Recognition and Emotion Analysis Based on Web Text Data. *Front. Psychol.* 12, 741928. doi:10.3389/fpsyg.2021.741928

Guo, X., Zhang, H., Yang, H., Xu, L., and Ye, Z. (2019). A Single Attention-Based Combination of CNN and RNN for Relation Classification. *IEEE Access* 7, 12467–12475. doi:10.1109/ACCESS.2019.2891770

Huang, Z., Wei, X., and Kai, Y. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. Chongqing: Computer Science. doi:10.48550/arXiv.1508.01991

Huiqin, W., and Weiguo, L. (2018). "Analysis of the Art of War of Sun Tzu by Text Mining Technology," in 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (IEEE). doi:10.1109/ICIS.2018.8466540

Kong, J., Zhang, L., Jiang, M., and Liu, T. (2021). Incorporating Multi-Level CNN and Attention Mechanism for Chinese Clinical Named Entity Recognition. *J. Biomed. Inf.* 116, 103737. doi:10.1016/j.jbi.2021.103737

Ma, X., and Hovy, E. (2016). *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*. IEEE. doi:10.13140/RG.2.1.2182.5685

Meng, L., and Xiang, J. (2018). Brain Network Analysis and Classification Based on Convolutional Neural Network. *Front. Comput. Neurosci.* 12, 95. doi:10.3389/fncom.2018.00095

Miwa, M., and Bansal, M. (2016). "End-to-End Relation Extraction Using LSTMs on Sequences and Tree Structures," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (IEEE). doi:10.18653/v1/P16-1105

Pal, N. R. (2020). In Search of Trustworthy and Transparent Intelligent Systems with Human-like Cognitive and Reasoning Capabilities. *Front. Robot. AI* 7, 76. doi:10.3389/frobt.2020.00076

Qiu, J., Zhou, Y., Wang, Q., Ruan, T., and Gao, J. (2019). Chinese Clinical Named Entity Recognition Using Residual Dilated Convolutional Neural Network with Conditional Random Field. *IEEE Trans.on Nanobioscience* 18 (3), 306–315. doi:10.1109/TNB.2019.2908678

Shu, Z., Dequan, D., and Xinchen, H. (2015). "Bidirectional Long Short-Term Memory Networks for Relation Classification," in The 29th Pacific Asia Conference on Language, Information and Computation (Shanghai: Scimago Lab).

Wickramarachchi, R., Henson, C., and Sheth, A. (2021). Knowledge-infused Learning for Entity Prediction in Driving Scenes. *Front. Big Data* 4, 759110. doi:10.3389/fdata.2021.759110

Wu, Q., Li, Q., Zhou, J., Long, X., Guan, Y., and Lin, H. (2020). "A Film and TV News Digest Generation Method Based on HanLP," in 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA) (IEEE). doi:10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00178

Yanchao, H., Yuanzhe, Z., and Kang, L. (2017). "An End-To-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (IEEE).