Journal of Nonlinear and Convex Analysis
Volume 26, Number 7, 2025, 2245–2259

Yokohama Publishers
ISSN 1880-5221 ONLINE JOURNAL

AN ENTITY RECOGNITION MODEL FOR THE COKING OF ETHYLENE CRACKING FURNACE BASED ON SELF-ATTENTION AND ROBERTA

JINGLONG ZUO, DELONG CUI*, ZHIPING PENG, QIRUI LI, JIEGUANG HE,
AND JIANBIN XIONG

Abstract. The entity recognition process of unstructured corpus text related to ethylene coking can be time-consuming and arduous. This study aimed to propose an entity recognition model based on self-attention and RoBERTa. First, corpus pretraining was achieved using the RoBERTa model, which could learn corpus with longer sequences and learn semantic expressions at the word level better compared with BERT (Bidirectional Encoder Representations from Transformers, BERT) BERT. Then, the preprocessed character feature sequence was subjected to BiLSTM (Bidirectional Long Short-Term Memory, BiLSTM) for semantic feature extraction. The self-attention mechanism was then integrated to perform secondary feature extraction on the character sequence to explore the correlation between entities. Finally, the prediction results were output using a conditional random field. The experimental results demonstrated that the proposed entity recognition model performed well in terms of accuracy, precision, and recall.

1. Introduction

Ethylene is one of the most fundamental raw materials used in the petrochemical industry. Hence its production is an essential criterion for judging the development level of a country's petrochemical industry. Currently, ethylene is mainly produced using coil cracking furnaces, which indicates that cracking furnaces are the core component of the refining system. Nevertheless, cracking the inner wall of the furnace coil can cause carbon buildup due to a pyrolysis reaction, resulting in coking and carburization. This phenomenon hinders the heat transfer of high-temperature flue gases to the raw materials inside the coil, thus reducing the production efficiency of ethylene. This incurs huge economic losses to enterprises and poses a safety hazard. Therefore, diagnosing the coal coking accurately during the production process is essential to increase the capacity of the ethylene cracking furnace.

2020 *Mathematics Subject Classification.* 68T07, 68U10.

Key words and phrases. Deep learning, entity recognition, ethylene cracking furnace, knowledge graph.

The work presented in this paper was supported by: National Natural Science Foundation of China (62273109); Key Realm R&D Program of Guangdong Province (2021B0707010003); Guangdong Basic and Applied Basic Research Foundation (2022A1515012022, 2023A1515240020, 2023A 1515011913, 2024A1515012090); Key Field Special Project of Department of Education of Guangdong Province (2024ZDZX1034); Maoming Science and Technology Project (210429094551175, 2022DZXHT028, mmkj2020033); Projects of PhDs' Start-up Research of GDUPT (2023bsqd2012, 2023bsqd1002, 2023bsqd1013, XJ2022000301).

*Corresponding author.

A knowledge graph has a strong ability to express things. It has been widely used in medical, chemical, and other industries. Segler [21] established a knowledge graph for binary chemical reactions considering the relationship between nodes in the graph. They deduced a new set of reaction equations through knowledge inference to compensate for the missing links in the knowledge graph. Mao [16] developed a safety knowledge graph for the delayed coking process using a combination of top-down and bottom-up approaches to define the process safety model at the ontology level. This approach enhanced the knowledge-based analysis capability, discovered the hidden relationships between the possible risk causes and consequences in emergency situations, and provided the basis for more process safety-related applications.

However, the entity recognition working process of unstructured corpus text related to ethylene coking requires operators to have knowledge reserves in various aspects, and is arduous and time-consuming. This study aimed to propose an entity recognition model based on self-attention and RoBERTa. First, the RoBERTa model pretrained the corpus relative to BERT (Bidirectional Encoder Representations from Transformers). The model could learn corpus with longer sequences and had a stronger learning ability for word-level semantic expression. The preprocessed character sequence was subjected to BiLSTM (Bidirectional Long Short-Term Memory) for semantic feature extraction, which was fused with a self-attention mechanism for secondary feature extraction, to explore the correlation between entities. Finally, the prediction results were output using a CRF (Conditional Random Field). The proposed entity recognition model had better performance in accuracy, precision, and recall.

The main contributions of this paper are as follows:

- We propose a self-attention-based RoBERTa-BiLSTM-CRF model for unstructured corpora.
- The model first pretrains the corpus using RoBERTa to obtain vector representations of the characters. These vector sequences are then used as input parameters to extract semantic features through the BiLSTM network model. Subsequently, the attention mechanism is employed to perform secondary feature extraction on the parameters. Finally, the CRF layer outputs the prediction results.
- The proposed model exhibits excellent performance in terms of accuracy, precision, and recall.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of the related work on ethylene cracking furnace coking research. Section 3 delves into the construction of the self-attention-based RoBERTa-BiLSTM-CRF text entity recognition model, along with a detailed analysis of the model's implementation. Section 4 presents the experimental results, and finally, Section 5 concludes the paper and outlines the framework for our future work.

2. Literature review

2.1. **Modeling of ethylene cracking.** Kumar [13] developed a naphtha molecular kinetic model based on seven types of pyrolysis raw materials to analyze the ethylene pyrolysis process. However, this model could not be directly applied to industrial production because it was highly dependent on process parameters not directly measurable. Moreover, the computational procedure was complex, and the calculation amount was too large. However, the ethylene pyrolysis mechanism could be roughly simulated and the time cost for the pyrolysis process calculation could be reduced after simplifying the complex mechanism and establishing 1D and 2D pyrolysis, coking kinetic model, chamber heat transfer model, and other process models for ethylene cracking furnace [5]. The regional method [8] was rapidly developed in the simulation field with the development of computer technology. Nevertheless, the model equations were extremely time-consuming and had low accuracy. Also, obtaining the temperature solution for the full process and full coil was difficult, making it hard to quickly reflect the change in the coil temperature with the turbulence of high-temperature flue gas in the chamber. CFD(Computational Fluid Dynamics) technology was widely used to simulate the production process of the cracking furnace. The raw materials entering the cracking furnace in the convection section were required to be vaporized beforehand [10], and those entering the radiation section were required to be preheated. The radiation section was the main reaction stage of raw materials, and CFD could simulate the turbulence of high-temperature flue gas in the chamber [18], thermal radiation and heat transfer through the tubes [17], and the effect of chamber geometry and dimensions, the structure of the combustor, and the distribution of the combustor [20] on the yield and quality of the pyrolysis product. However, accurate CFD simulations require a large number of complex iterative calculations and are extremely dependent on the accuracy of process parameters, which are often difficult to obtain.

2.2. **Application of neural networks in the cracking furnace.** The deep learning algorithm showed strong nonlinear function approximation capability with the rapid increase in computational power. Hence, a deep learning algorithm is gradually being applied to chemical process simulation, product yield prediction, coil coking prediction, and so on. The obtained coil temperatures are often incorrect due to the spatial arrangement of the coils and the overlapping temperatures of the coils. Zhao et al. [31] proposed an intelligent temperature measurement device designed to measure the coil temperature. The device uses machine learning and CNN (Convolutional Neural Networks) neural networks to identify the overlapping coil temperature data in the measurements. This technology, to a certain extent, not only reduced the cost of the coil temperature measurement but also improved the accuracy of the measured data. Xia et al. [25] identified a nonlinear multivariate system using RBFNN and proposed a fuzzy C mean multiswarm competitive particle swarm (FCMCPSO) algorithm for optimizing controlled variables of a real-time computing system. This algorithm could effectively control the depth of ethylene pyrolysis in the ethylene plant and improve the yield of ethylene and propylene. The ethylene cracking furnace chamber temperature is one of the essential indicators of the health degree of the ethylene cracking furnace, and the abnormal temperature data directly affects the stability and reliability of the ethylene cracking furnace production. Chen et al. [4] established a prediction algorithm and model for the trend of ethylene cracking furnace chamber temperature data based on the Bayesian optimized SVM (Support Vector Machine) regression.

They compared the prediction algorithm with linear regression and the random forest algorithm to validate the effectiveness of the model. This method could well predict the variation trend of the ethylene cracking furnace chamber temperature, allowing for timely monitoring the working status of the furnace. Hua et al. [11] established a pyrolysis knowledge graph based on the naphtha pyrolysis knowledge, and developed a novel naphtha pyrolysis model on the basis of structural features in the CNN learning response knowledge graph. This model had a better learning effect for kinetic, and the computational cost was lower; it could predict the yield of key products with high accuracy compared with conventional naphtha mechanism models. Xin et al. [23] compared the generalized regression neural networks model and backpropagation neural network model for coke yield prediction with industrial production data. They further optimized the BP (Back Propagation) neural networks for accuracy and stability using a particle swarm optimization algorithm to further improve model accuracy. Liu et al. [14] proposed an Adaboost-based ethylene pyrolysis coking hybrid prediction model. They employed the Adaboost algorithm to focus on learning the features of misclassified samples so as to improve the model accuracy.

2.3. **Diagnosis of coil coking of the ethylene cracking furnace.** The diagnosis of coil coking of the ethylene cracking furnace was thoroughly investigated. The first kind involved the construction of the coking model. For different pyrolysis raw materials, a mathematical analytical equation was developed as a coking model, with pyrolysis process parameters as inputs and coking rate and slag thickness as outputs [1][7][12][22][24][29]. The relationship between the petroleum fraction and the generated gas in the pyrolysis process was analyzed, and a liquid-gas two-phase CFD method was proposed to predict the light petroleum fraction, the generated gas composition, and the thickness of the coke generated in the tube during the pyrolysis process [24]. Xx et al. [22] analyzed the laminar and turbulent flows of propane and naphtha fluids, and developed a dynamic mathematical model based on the second-order turbulence model to predict the coke generation of propane and naphtha fluids in coils. A dynamic ethylene cracking furnace coil coking model was developed based on the pyrolysis reaction process to predict the slag thickness under various operating conditions [12]. Such methods usually require a combination of experimental or industrial data to determine the coking kinetics and adjustable parameters in the coking model. Additionally, some key model parameters, including activation energy and frequency factor, can barely be measured. Hence, the accuracy of coking inference based on the coking model is generally not high. The second kind involved data-driven intelligent diagnosis. Data-driven methods have gained great attention with the development of AI (Artificial Intelligence) and Big Data technologies. Usually, the sensible variables related to coking were selected as inputs, and AI algorithms were designed to construct a “black box” model to obtain the relationship between slag thickness and measurable input variables. Su et al. [23] used a backpropagation neural network model optimized by a genetic algorithm to predict the production rate of coke in a catalytic pyrolysis plant. Chen et al. [3] used SVM to identify the working conditions and develop a stochastic distribution system model for the COT (Coil Outlet Temperature) in the ethylene cracking furnace burning process. They aimed to lay the foundation for advanced stochastic distribution control of the COT in the ethylene cracking furnace burning process. The project team proposed a diagnosis and prediction method for the coking of an ethylene cracking furnace by integrating an artificial bee colony algorithm and an adaptive neural fuzzy inference system [19]. Peng et al. [2] proposed a pyrolysis optimization model combining transfer learning and heuristic algorithms. First, the features of the slag thickness model and the product yield prediction model were obtained through transfer learning. Then, the slag thickness and product yield were predicted using heuristic optimization algorithms.

Overall, AI technologies, such as neural networks and transfer learning, have been used for diagnosing coil coking in ethylene cracking furnaces, achieving great advances. Nevertheless, current intelligent diagnosis techniques focused on qualitative analysis, while few focused on quantitative analysis. Additionally, most methods are data-driven, resulting in poor robustness and interpretability.

3. Self-attention-based RoBERTa-BiLSTM-CRF text entity recognition model

A self-attention-based RoBERTa–BiLSTM–CRF model for the unstructured corpus associated with ethylene coking was proposed, as shown in Figure 1. First, the corpus was pretrained by RoBERTa [15] to obtain the vector features of the characters. Then, the obtained vector sequences were used as input parameters to extract semantic features through the BiLSTM network model. The parameters were subjected to secondary feature extraction by the attention mechanism. In addition, the prediction results were output from the CRF.

3.1. **Data pretreatment.** To date, BIO [9] and BIOES [26] are the most common approaches for sequence labeling. BIO includes three labels: “B-X” indicates the word is at the beginning of an entity, “I-X” indicates the word is inside an entity but not at the beginning, and “O” indicates the word is outside any entity.

BIOES extends BIO by adding “E-X” to indicate the word is at the end of an entity and “S” to indicate the word is a single-word entity. “B-X”, “I-X”, and “O” in BIOES have the same meaning as in the BIO annotation.

3.2. **RoBERTa pretraining layer.** Similar to BERT [6], RoBERTa is also composed of stacked transformer structures and is trained on a large amount of text data. However, unlike BERT, it eliminates next sentence prediction tasks and can handle larger batches of data. More importantly, it changes the BERT static masking to a dynamic masking strategy, which includes copying the data to be trained into 10 copies and randomly selecting 15% of each sequence to be dynamically masked, that is, dynamically changing the masking of each input sequence.

BERT learns a priori semantic representations of words by statically masking the words of a sentence, but it only learns word-level feature information. However, RoBERTa can learn more semantic representations of characters and, at the same time, learn semantic representations of words by masking the same corpus with

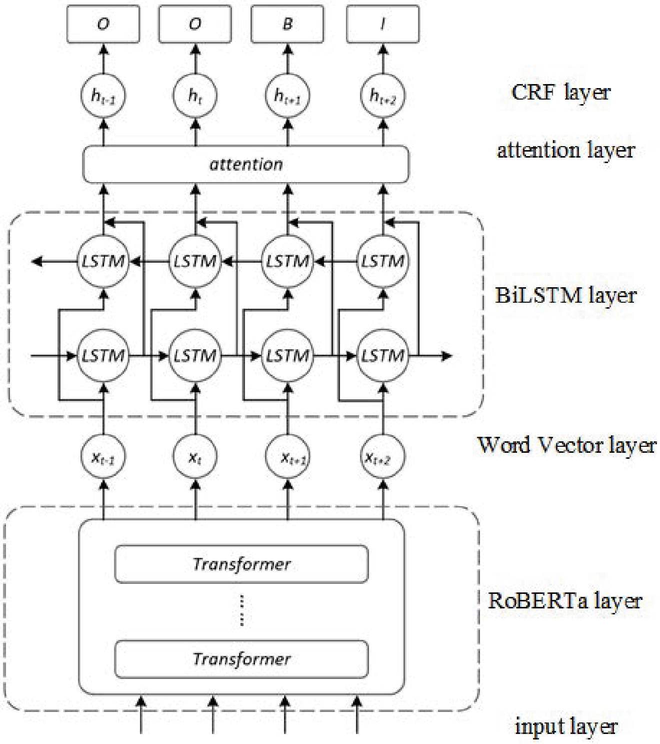

FIGURE 1. Relationship schema and mapping attributes

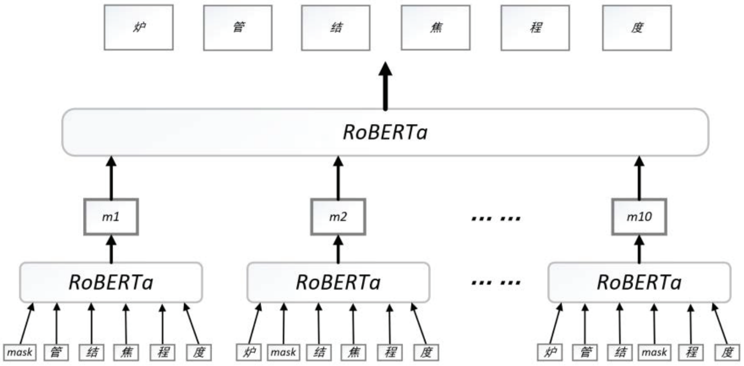

FIGURE 2. Pretreatment by RoBERTa.

random dynamic characters during the pretraining of the Chinese corpus. Therefore, this model has a stronger semantic learning ability for Chinese characters. Pretreatment by RoBERTa is illustrated in Figure 2.

3.3. **BiLSTM network layer.** LSTM (Long Short-Term Memory, LSTM) introduces gating units to control the update of cell states and the flow of information, including forget gates, input gates, and output gates. The forget gate determines

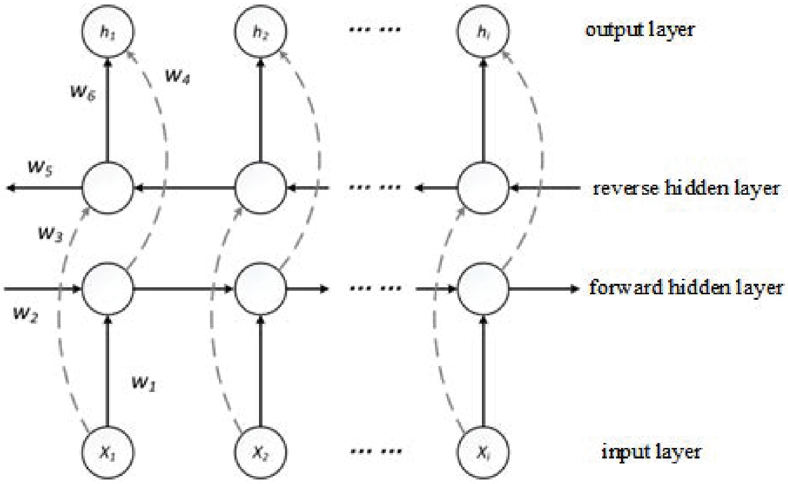

Figure 3. Structure of BiLSTM.

which information from the previous cell state is retained; the input gate controls which current input information is written into the cell state; and the output gate determines which information from the cell state is output at the current step.

Through these gating mechanisms, LSTM can address the long-term dependency issue of traditional RNNs to a certain extent, enabling it to better handle long-distance dependencies in sequential data. However, LSTM is a unidirectional network structure that uses past information to predict future information, but is less effective in predicting past information from future information. To address this issue, the network model used in this paper is BiLSTM, the network structure is shown in Figure 3. BiLSTM consists of forward and reverse LSTMs, comprising the input layer, forward hidden layer, reverse hidden layer, and output layer. The input sequences $x = x_1, x_2, ..., x_n$ are shared by forward and reverse LSTMs, which realize the purpose of focusing on the context information of the input sequences at the same time.

$$h_{t,f} = LSTM(x_t, h_{t-1,f}), \tag{3.1}$$

$$h_{t,b} = LSTM(x_t, h_{t-1,b}) \tag{3.2}$$

where x_t refers to the input information at moment t; and $h_{t,f}$ and $h_{t,b}$ are the information of forward and reverse hidden layers, respectively, at moment t.

By learning features from the input sequences, forward and reverse LSTMs each output a sequence of feature information $h_{t,f}$ and $h_{t,f}$, fusing forward and reverse feature sequences to obtain the final network output $y = y_1, y_2, ..., y_n$:

$$y_t = f(W_f h_{t,f} + W_b h_{t-1,b} + b) \tag{3.3}$$

where W refers to the weight matrix of the network, and b refers to the bias.

At each time step, BiLSTM concatenates or fuses the forward hidden state and the backward hidden state to obtain the final hidden state representation for that time step. This hidden state representation thus incorporates information from both the forward and backward directions of the sequence, enabling it to more comprehensively capture the sequence's features and contextual relationships.

3.4. **Attention mechanism.** The entities in a sentence are related to each other to a certain extent, but some entities are distributed far away. Meanwhile, as the distance between entities increases, the ability of LSTM networks to obtain the relationship between these entities decreases. The correlation between entities can be directly evaluated without considering their distances using the attention mechanism. Hence, in entity recognition, the discrimination model should focus more on the feature information of the current character, increase the weight of the character with a strong correlation, decrease the weight of the character with a weak correlation, and finally use the feature information of the character to improve the performance of the discrimination model.

This study employed the self-attention mechanism, which could capture associative relationships between words regardless of the distance between them. First, the input vector x was multiplied by the attention score matrices W_k and W_q to obtain two scores $f(x)$ and $g(x)$ of the vector, respectively. Then, the vector's dimension was divided by the matrix product to obtain the vector's attention distribution. Finally, the attention matrix s was normalized by the softmax function.

$$f(x) = W_k x, g(x) = W_q x, \tag{3.4}$$

$$s(x) = \frac{f(x)^T g(x)}{\sqrt{d}}, \tag{3.5}$$

$$a_i = softmax(s(x)) = \frac{\exp(s(x))}{\sum_{j=1}^{n} \exp(s(x))}. \tag{3.6}$$

The eventual output of the attention layer was obtained by numerically summing the normalized attention distribution matrix a_i and input vector x:

$$y'' = \sum_{i=1}^{n} a_i \cdot x. \tag{3.7}$$

Each element in the output sequence is a weighted sum of all value vectors, effectively integrating information within the sequence such that the output representation of each element encapsulates the contextual information of the entire sequence.

3.5. **CRF layer.** CRF is a probabilistic graphical model used for modeling sequential data. The fundamental idea behind CRF is that, given a set of input variables (such as word sequences, image pixels, etc.), the joint probability distribution of the output variables (such as label sequences, image segmentations, etc.) can be represented by an undirected graph, where each node corresponds to a variable and each edge corresponds to a potential function that measures the correlation between variables. The goal of CRF is to learn the structure and parameters of this undirected graph, enabling optimal prediction of output variables for new input variables.

In this paper, the label prediction was performed using the CRF model, which took into account the global information of the label sequence. The parameter of the CRF layer was a transfer matrix A, where $A_{i,j}$ denotes the transfer score from the i_{th} label to the j_{th} label, which can be used to mark a position using a previously labeled label. Let the sentence length be n and the output label sequence $y = (y_1, y_2, ..., y_n)$, then the prediction score of this label sequence was obtained as:

$$S(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}. \tag{3.8}$$

The conditional probability of sequence y over all possible sequences could be calculated using the softmax function:

$$P(y, X) = \frac{e^{S(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{S(X,y)}}. \tag{3.9}$$

During the model training process, the probability P needed to be transformed into a log function:

$$log(P(y, X)) = s(X \mid y) - log(\sum_{\tilde{Y} \in Y_X} e^{s(X, \tilde{Y})}). \tag{3.10}$$

Sentence X, Y represent all possible label sequences. The labeling sequence with the highest conditional probability was obtained by decoding:

$$y^* = \underset{\tilde{y} \in Y_X}{argmaxs}(X, \tilde{y}). \tag{3.11}$$

4. Results and Discussion

4.1. **Datasets.** In the proposed ethylene coking knowledge graph, the corpus was sourced from various avenues, including operational information pertaining to ethylene production, details of ethylene producers from a petrochemical company in China, published books related to ethylene technology and coking mechanisms, as well as relevant literature retrieved from CNKI. The corpus comprises text instances that delineate various facets of ethylene cracking furnace operations, such as types of raw materials, process conditions, coking phenomena, and maintenance records.

Each data instance within the corpus was meticulously annotated with entity labels utilizing the BIOES scheme. For instance, the sentence "the ethylene cracking furnace uses naphtha as the raw material". might be annotated as "The ethylene [B-Entity] cracking [I-Entity] furnace [I-Entity] uses [O] naphtha [B-Entity] as [O] the [O] raw [O] material [O]". Here, "B-Entity" signifies the onset of an entity, "I-Entity" denotes the continuation of an entity, and "O" indicates that the token is not a part of any entity.

The annotated corpus was subsequently divided into training, validation, and test sets, with respective ratios of 8:1:1. The training set was employed to train the proposed model, the validation set was utilized to fine-tune hyperparameters, and the test set was used to assess the model's performance.

4.2. **Indicators.** Accuracy (Acc), precision (Pre), recall (Rec), and F1 score (F1) were employed as indicators in this study. The calculation results were recorded in a confusion matrix, which contained information about the actual classification of the data and the model's predicted classification. In the confusion matrix, TP denotes the number of samples in which the model classified the actual normal data as normal, FN denotes the number of samples in which the model classified the actual normal data as abnormal, FP denotes the number of samples in which the model classified the actual abnormal data as normal, and TN denotes the number of samples in which the model classified the actual abnormal data as abnormal.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.1}$$

$$Pre = \frac{TP}{TP + FP}, \tag{4.2}$$

$$Rec = \frac{TP}{TP + FN}, \tag{4.3}$$

$$F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}. \tag{4.4}$$

4.3. **Results.** The annotated corpus was used to train the proposed model. During the training process, each data instance in the training set was processed as follows: - The input sentence was tokenized and converted into a sequence of character embeddings using the RoBERTa pretraining layer. The character embeddings were then fed into the BiLSTM network layer to extract contextual semantic features. The self-attention mechanism was applied to the BiLSTM output to capture the correlations between entities regardless of their distances.

Finally, the CRF layer was used to decode the sequence of labels, taking into account the global information of the label sequence. The model was trained using the cross-entropy loss function, which measures the difference between the predicted label sequence and the true label sequence. The model parameters were optimized using the Adam optimizer. Figure 4 shows the results obtained using BIO and BIOES sequence labeling methods. Compared with the BIO labeling, the BIOES labeling had certain advantages in all three indicators, wherein the precision, recall, and F1 score increased by 1.45%, 1.27%, and 1.36, respectively. Additionally, the BIOES labeling could provide more segmentation information than the BIO labeling, resulting in enhanced recognition efficiency. This study employed BIOES labeling for the sequence labeling model.

ELMo-MT-BBC [28] is an earning model based on multi-task attention. This model improves the prediction performance by incorporating a pretrained semantic model, reduces the noise in entity recognition using a regularization mechanism, and builds a multi-tasking mechanism to enhance the model's perception of unknown entities to improve the recall.

Semi-BBLC [30] is a semi-supervising and embedded entity recognition model. A small amount of labeled data is used to train the model, the model is used to add pseudo-labeling to the unlabeled data, and finally all the data are used to continue to train a discriminative model with strong recognition ability.

XL-BC [27] is an XLNet-based entity recognition model. The XLNet, as a pretraining model, solves the problem that word vectors cannot be accurately recognized from small datasets.

Table 1 summarizes the experimental results of different models. The accuracy, precision, recall, and F1 score of the proposed model were 75.61%, 90.91%, 79.37%,

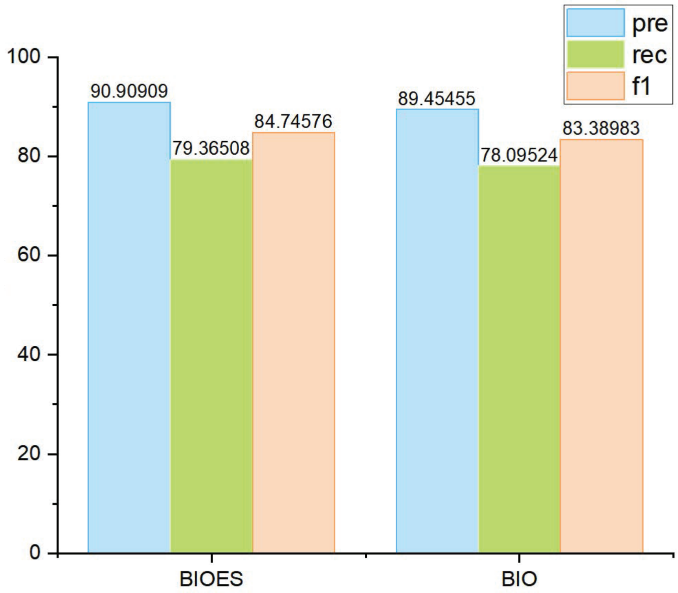

FIGURE 4. Sequence labeling results of BIO and BIOES.

and 84.75%, respectively. The F1 score of the proposed model was 1.69%, 3.11%, and 4.47% higher than that of ELMo-MT-BBC, XL-BC, and semi-BBLC models, respectively. The recall of the proposed model was 1.59%, 3.17%, and 6.35% higher than that of ELMo-MT-BBC, XL-BC, and semi-BBLC models, respectively. The precision of the proposed model was 1.82%, 3.0%, and 1.76% higher than that of ELMo-MT-BBC, XL-BC, and semi-BBLC models, respectively. The accuracy of the proposed model was 2.71%, 4.88%, and 6.23% higher than that of ELMo-MT-BBC, XL-BC, and semi-BBLC models, respectively. Overall, the proposed model exhibited advantages in all indicators involved.

TABLE 1. Experimental comparison results of the model

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
Semi-BBLC	69.38	89.15	73.02	80.28
XL-BC	70.73	87.91	76.19	81.63
ELMo-MT-BBC	72.90	89.09	77.78	83.05
The proposed model	75.61	90.91	79.37	84.75

Figure 5 shows the changes in the indicators of the proposed model during the training process. Before the 16th round of training, all the indexes kept increasing rapidly; in the 20th round of training, all the indexes entered into the stage of slow increase in value; and in the 73rd round of training, all the indexes were close to the stable and unchanged state. Therefore, the results of the 73rd round of training were taken as the final experimental results.

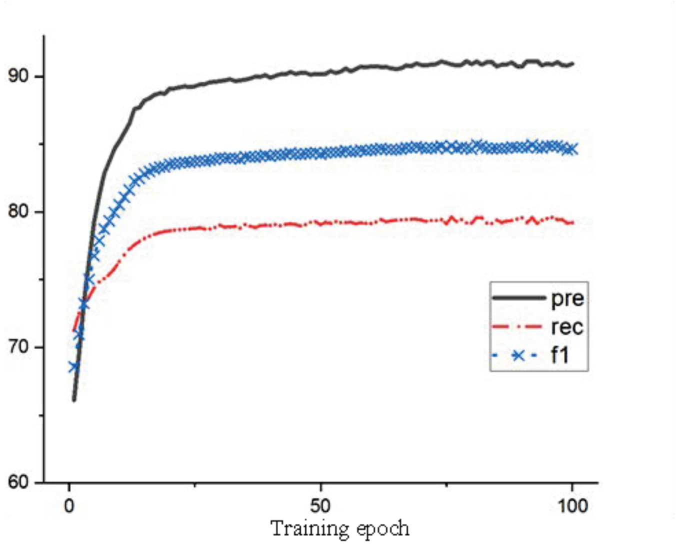

Figure 5. Changes in each index of the model.

5. Conclusions

An novel self-attention-based RoBERTa-BiLSTM-CRF model for entity recognition in unstructured corpora related to ethylene coking is proposed. Our work significantly contributes to the general knowledge body of current research in the literature by introducing a novel approach that leverages advanced deep learning techniques to tackle the challenging task of entity recognition in the context of ethylene cracking furnaces.

The model exhibits superior performance in terms of accuracy, precision, and recall, as evidenced by our experimental results. The significance, originality, and contribution of our study lie in its ability to effectively handle the complexities and ambiguities inherent in unstructured texts related to ethylene coking, thereby filling a critical research gap in this field. The key outcomes of our research include the development of a robust entity recognition model and the validation of its effectiveness through rigorous experimentation. Our findings have the potential to direct stakeholders and the research community towards more accurate and efficient diagnosis of ethylene cracking furnace coking, ultimately enhancing production efficiency and safety.

In considering the methodology followed in our analysis, we acknowledge that certain limitations exist. One such limitation is the reliance on a specific dataset for model training and testing, which may limit the generalizability of our results. Additionally, while our model performs well on the given tasks, there is always room for improvement in terms of accuracy and efficiency. To address these limitations, we propose several improvement points. Firstly, we plan to expand our dataset to include more diverse sources of information, thereby enhancing the generalizability of our model. Secondly, we aim to incorporate more advanced deep learning techniques to further improve the accuracy and efficiency of our entity recognition model. These improvement points will guide future research in this field, enabling other researchers to build upon our work and develop even more sophisticated models for entity recognition in unstructured corpora related to ethylene coking.

In future work, we will endeavor to extend the application of our self-attention-based RoBERTa-BiLSTM-CRF model to a broader spectrum of domains within the petrochemical industry, with a particular emphasis on enhancing its generalizability and robustness. We aim to integrate domain-specific knowledge and data to further refine the model's accuracy and precision in recognizing entities pertinent to diverse cracking processes and equipment fault diagnoses. Furthermore, we plan to explore the integration of our model with real-time monitoring systems, thereby enabling dynamic and predictive maintenance of ethylene cracking furnaces.

References

[1] A. Barza, B. Mehri and V. Pirouzfar, *Mathematical modeling of ethane cracking furnace of olefin plant with coke formation approach*, International Journal of Chemical Reactor Engineering **16** (2018): 20170243.

[2] K. Bi, B. Beykal, S. Avraamidou, I. Pappas, E N. Pistikopoulos and T. Qiu, *Integrated modeling of transfer learning and intelligent heuristic optimization for a steam cracking process*, Industrial & Engineering Chemistry Research **59** (2020), 16357–16367.

[3] D. Chen, X. Wan and Z. Wan, *Modeling of COT random distribution system for decoking process of ethylene cracking furnace based on operating region recognition*, Chemical Automation and Instrumentation **49** (2022), 47–53.

[4] W. Chen, S. Hu and H. Song, *Temperature data prediction and anomaly detection of ethylene cracking furnace based on SVR*, in: Proceedings of 2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS). IEEE. 2021, pp. 1–5.

[5] S. Dai, Y. Liang, S. Liu, Y. Wang, W. Shao, X. Lin and X. Feng, *Learning entity and relation embeddings for knowledge graph completion*, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence. AAAI Press, 2015, pp. 2181–2187.

[6] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li and X. Bai, *Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records*, in: Proceedings of 2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei). IEEE. 2019, pp. 1–5.

[7] A. Davarpanah, M. Zarei, K. Valizadeh and B. Mirshekari, *CFD design and simulation of ethylene dichloride (EDC) thermal cracking reactor*, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects **41** (2019), 1573–1587.

[8] Z. Fang, T. Qiu and W. Zhou, *Coupled simulation of recirculation zonal firebox model and detailed kinetic reactor model in an industrial ethylene cracking furnace*, Chinese Journal of Chemical Engineering **25** (2017), 1091–1100.

[9] L. He, K. Lee, O. Levy and L. Zettlemoyer, *Jointly predicting predicates and arguments in neural semantic role labeling*, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2018, pp. 364–369.

[10] G. Hu, J. Long and W. Du, *Numerical simulation of convection section of ethylene cracking furnace considering evaporation effect*, Journal of East China University of Science and Technology **45** (2019), 719–727.

[11] F. Hua, Z. Fang and T. Qiu, *Modeling ethylene cracking process by learning convolutional neural networks*, Computer Aided Chemical Engineering **44** (2018), 841–846.

[12] H. Karimi, B. Olayiwola, H. Farag and K B. McAuley, *Modelling coke formation in an industrial ethane-cracking furnace for ethylene production*, The Canadian Journal of Chemical Engineering **98** (2020), 158–171.

[13] P. Kumar and D. Kunzru, *Modeling of naphtha pyrolysis*, Industrial & Engineering Chemistry Process Design and Development **24** (1985), 774–782.

[14] L. liu, Y. Li and L. Fang, *Soft measurement of ethylene cracking deposit quantity based on adaboost hybrid model*, China Automation and Instrumentation No. 06 (2015), 50–53.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, CoRR. 2019, DOI: 10.48550/arXiv. 1907. 11692.

[16] S. Mao, Y. Zhao, J. Chen, B. Wang and Y. Tang, *Development of process safety knowledge graph: A Case study on delayed coking process*, Computers & Chemical Engineering **143** (2020): 107094.

[17] P. Mu, X. Gu and Q. Zhu, *Modeling and optimization of ethylene cracking feedstock scheduling based on P-graph*, CIESC Journal **70** (2019), 556–563.

[18] C. Ni, W. Du and G. Hu, *Impact of turbulence model in coupled simulation of ethylene cracking furnace*, CIESC Journal **70** (2019), 450–459.

[19] Z. Peng, J. Zhao, Z. Yin, Y. Gu, J. Qiu and D. Cui, *ABC-ANFIS-CTF: A Method for diagnosis and prediction of coking degree of ethylene cracking furnace tube*, Processes **7** (2019): 909.

[20] J. G. Rebordinos, C. Herce, A. Gonzalez-Espinosa, M. Gil, C. Cortés, F. Brunet, L. Ferré and A. Arias, *Evaluation of retrofitting of an industrial steam cracking furnace by means of CFD simulations*, Applied Thermal Engineering **162** (2019): 114206.

[21] M. Segler and M. Waller, *Modelling chemical reasoning to predict and invent reactions*, Chemistry-A European Journal **23** (2017), 6118–6128.

[22] A. R. Solaimany Nazar, F. Banisharifdehkordi and S. Ahmadzadeh, *Mathematical modeling of coke formation and deposition due to thermal cracking of petroleum fluids*, Chemical Engineering & Technology **39** (2016), 311–321.

[23] X. Su, Y. Wu, H. Pei, J. Gao and X. Lan, *Prediction of Coke yield of FCC unit using different artificial neural network models*, China Petroleum Processing & Petrochemical Technology **18** (2016), 102–109.

[24] M. G. Valus, D. V. R. Fontoura, R. Serfaty and J. R. Nunhez, *Computational fluid dynamic model for the estimation of coke formation and gas generation inside petrochemical furnace pipes with the use of a kinetic net*, The Canadian Journal of Chemical Engineering **95** (2017), 2286–2292.

[25] L. Xia, J. Chu and Z. Geng, *A multiswarm competitive particle swarm algorithm for optimization control of an ethylene cracking furnace*, Applied Artificial Intelligence **28** (2014), 30–46.

[26] L. Yan and S. Li, *Grape diseases and pests named entity recognition based on BiLSTM-CRF*, in: Proceedings of 2021 IEEE 4th Advanced Information Management. Communicates, Electronic and Automation Control Conference (IMCEC), IEEE. 2021, pp. 2121–2125.

[27] R. Yan, X. Jiang and D. Dang, *Named entity recognition by using XLNet-BiLSTM-CRF*, Neural Processing Letters **53** (2021), 3339–3356.

[28] J. Yang, Y. Liu, M. Qian, C. Guan and X. Yuan, *Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding*, Applied Sciences **9** (2019): 3658.

[29] S. M. Zaker Abbasali, M. Farsi and M. R. Rahimpour, *Simulation and dynamic optimization of an industrial naphtha thermal cracking furnace based on time variant feeding policy*, Chemical Product and Process Modeling **13** (2018): 20170032.

[30] M. Zhang, Z. Yang, C. Liu and L. Fang, *Traditional Chinese medicine knowledge service based on semi-supervised BERT-BiLSTM-CRF model*, in: Proceedings of 2020 International Conference on Service Science (ICSS), IEEE, 2020, pp. 64–69.

[31] J. Zhao, Z. Peng, D. Cui, Q. Li, J. He and J. Qiu, *A method for measuring tube metal temperature of ethylene cracking furnace tubes based on machine learning and neural network*, IEEE Access **7** (2019), 158643–158654.

Manuscript received October 24, 2024
revised February 27, 2025

J. L. Zuo
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, School of Electronic Information Engineering, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China
E-mail address: oklong@foxmail.com

D. L. Cui
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, School of Electronic Information Engineering, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China
E-mail address: delongcui@gdupt.edu.cn

Z. P. Peng
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, Jiangmen Polytechnic, Jiangmen, Guangdong, China
E-mail address: zhipingpeng@gdupt.edu.cn

Q. R. Li
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, School of Computer Science and Engineering, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China
E-mail address: liqirui@gdupt.edu.cn

J. G. He
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, School of Computer Science and Engineering, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China
E-mail address: jieguanghe@gdupt.edu.cn

J. B. Xiong
Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, School of Automation, Guangdong Polytechnic Normal University, Guangdong, Guangzhou, China
E-mail address: 276158903@qq.com