


Research Article

Named Entity Recognition of Hazardous Chemical Risk Information Based on Multihead Self-Attention Mechanism and BERT

Guanlin Chen,^{1,2} Zhao Cheng,^{1,2} Qi Lu,³ Wenyong Weng,¹ and Wujian Yang¹ 

¹School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China

²School of Information Science & Engineering, Changzhou University, Changzhou 213164, China

³China National Air Separation Engineering Co., Ltd., Hangzhou 310051, China

Correspondence should be addressed to Wujian Yang; yangwj@zucc.edu.cn

Received 13 May 2022; Revised 6 June 2022; Accepted 22 June 2022; Published 7 July 2022

Academic Editor: Kuruva Lakshmana

Copyright © 2022 Guanlin Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

An approach based on self-attention mechanism and pretrained model BERT is proposed to solve the problems of entity recognition and relationship recognition of hazardous chemical risk information. The text of hazardous chemical risk information is coded at the character level by adding the pretrained language model BERT, which, when paired with a multihead self-attention mechanism, improves the ability to mine global and local aspects of texts. The experimental results show that the model's F1 value is 94.57 percent, which is significantly higher than that of other standard models.

1. Introduction

Human knowledge learning is one of the directions of artificial intelligence study. Knowledge representation and reasoning inspired by human problem-solving behavior enable the intelligent system to accurately describe and classify the acquired knowledge information and to obtain the ability to solve complex problems. The knowledge graph, a structured form of anthropological knowledge introduced by Google in 2012, has piqued academic and industry interest. Named entity recognition plays a key role in the process of constructing knowledge graphs.

Deep learning has advanced tremendously as hardware has improved. Compared with traditional methods, deep learning shows significant advantages in many tasks [1]. Mainstream named entity recognition approaches are built on deep learning, which converts the process of named entity recognition into a sequence annotation problem to fulfill the entity label prediction. With the proposal of BERT [2], the recognition performance of named entity recognition has been greatly improved. At the same time, the semantic information between sentences is mined by using

multihead self-attention, which makes the model obtain more features and improves the recognition performance to a certain extent. However, the self-attention mechanism in front of the BiLSTM layer is then used in many studies to obtain the characteristics of the original sentences. Since BERT also uses the self-attention mechanism, the effect of this structure is not very good. Furthermore, there is no relevant research on the named entity recognition of hazardous chemical risk information. Due to the lack of label data of hazardous chemical risk information, the effect of applying existing methods directly to the hazardous chemical risk information is not ideal, and the entity recognition accuracy is low.

In order to solve these problems, we first crawl and filter the web information to form a dataset, then design an annotation tool that conforms to the annotation characteristics of the dataset to label the dataset, and finally, this work suggests the BERT-BiLSTM-Self-Attention-CRF model, which will aid in the creation of a hazardous chemical knowledge graph. The text character level coding in the field of hazardous chemical risk information is obtained using the BiLSTM-CRF model, and the word embedding based on

context information is obtained using the pretrained language model BERT. The self-attention layer is added after the BiLSTM layer to mine the features of the output vector of the BiLSTM layer, which enhances the ability of the model to mine the global features between texts.

2. Related Works

The goal of the named entity recognition task is to extract entities from the target text that adhere to preset semantic kinds. It is a basic job in natural language processing that offers support for later activities. Currently, there are four main ways for recognizing named entities: (1) the rule-based method does not need to label data, but relies on manual rules. Based on rules, corresponding dictionaries can be made to improve the recognition effect. However, in the case of domain-specific rules or incomplete domain dictionaries, these systems usually have high accuracy and low recall rate, and these systems cannot be transferred to other fields. (2) The unsupervised learning method relies on unsupervised learning and does not need manual marking data. The typical method is clustering, which extracts named entities from the cluster group through the similarity of context. (3) Feature based on the statistical machine learning method relies on feature selection. Using supervised learning, the named entity recognition task is transformed into a multi-classification or sequence marking task. To acquire the target model, it is important to choose features from the text for annotation and train the features using a machine learning method. (4) The method based on deep learning is also the most used.

Compared with feature-based methods, deep learning algorithms can find hidden features more automatically, and the effect is better. In 1997, a long short-term memory network model (LSTM) was introduced, which may be selected to ignore the unnecessary information in the preceding text and increase named entity recognition accuracy. To address the issue that the LSTM model only uses prior timing information to forecast output at a future time without taking into account the next state, Strubell et al. [3] integrated the iterated dilated convolutional neural network (IDCNN) and conditional random field (CRF) architectures; Young et al. [4] designed combining bidirectional long short-term memory (BiLSTM) with conditional random field (CRF). To get richer general semantic expressions, BERT was proposed by Devlin et al. The pretrained word embedding with good representation capacity is obtained by training a large-scale corpus. To strengthen the ability of the model to obtain semantic information in the field, Xie et al. [5] increased entity recognition accuracy using the BiLSTM-CRF architecture and BERT model. However, the LSTM unit is limited to local information, and its ability to obtain long-distance information is insufficient, resulting in the deviation of the overall effect of the model.

In many industries, such as cybersecurity [6], biomedicine [7, 8], and social media [9, 10], entity naming recognition based on deep learning algorithms is frequently employed. Nevertheless, there is no relevant research on named entity recognition of hazardous chemical risk infor-

mation by scholars, and there is no large-scale labeled dataset in the field of hazardous chemical risk information. There are still some difficult problems in the entity identification method for hazardous chemical risk information, such as much data in the field of hazardous chemical risk information, different storage formats, and great individual differences. Information entities are not in a sentence in various data, and entities overlap with entity types in other fields and are inconsistent with common named entities. Conceptual entities have to be redefined. Hazardous chemical risk information is lacking due to a lack of labeling data, the high expense of manual labeling, the time and energy required, and the necessity for experienced employees to aid in labeling, making it difficult to label entities. As a result, this paper constructed a hazardous chemical risk information data set by crawling and filtering web information. And then the BERT-BiLSTM-Self-Attention-CRF model is proposed. By connecting the self-attention mechanism layer after the BiLSTM layer, the model can strengthen the mining ability of the global information in the statement, so that the model can be more effective in processing the risk information dataset of hazardous chemicals.

3. The Structure of BERT-BiLSTM-Self-Attention-CRF

The model proposed in this paper is based on BiLSTM-CRF, which replaces the input of BiLSTM from word2vec pretrained word embedding with BERT pretrained word embedding to produce a more informative word vector; at the same time, the self-attention mechanism layer is connected behind the BiLSTM layer to mine the semantic information between characters at a deeper level.

The characters are first fed into the BERT, which generates a word vector by combining word embedding, segment embedding, and position embedding. Then, the word vector fused with semantics is applied to the input of the BiLSTM network. Through BiLSTM, the model can learn how to predict the output of the next time through the timing information before and after. The self-attention mechanism layer is then used to mine the local features mined by the BiLSTM network, retrieve the output feature vector's interaction relationship, enhance the feature vector's global features, and supplement the features of the BiLSTM layer's output vector. Finally, the CRF layer learns the rules following the interaction between tags for the tag prediction variables produced from the BiLSTM layer, such as not linking B-SUBJECT after I-SUBJECT, to increase the logic of the prediction tag and allow the model to achieve the optimum output tag sequence. Figure 1 depicts the model structure.

3.1. Word Embedding Based on BERT. In natural language processing tasks, the frequently used word embedding models include Word2Vec proposed by Google and Glove proposed by Stanford University. Word2vec retrieves the associated word vector of the word based on its context; Glove enriches the semantic information of the word vector by using a co-occurrence matrix and considering local and overall information at the same time. However, these word

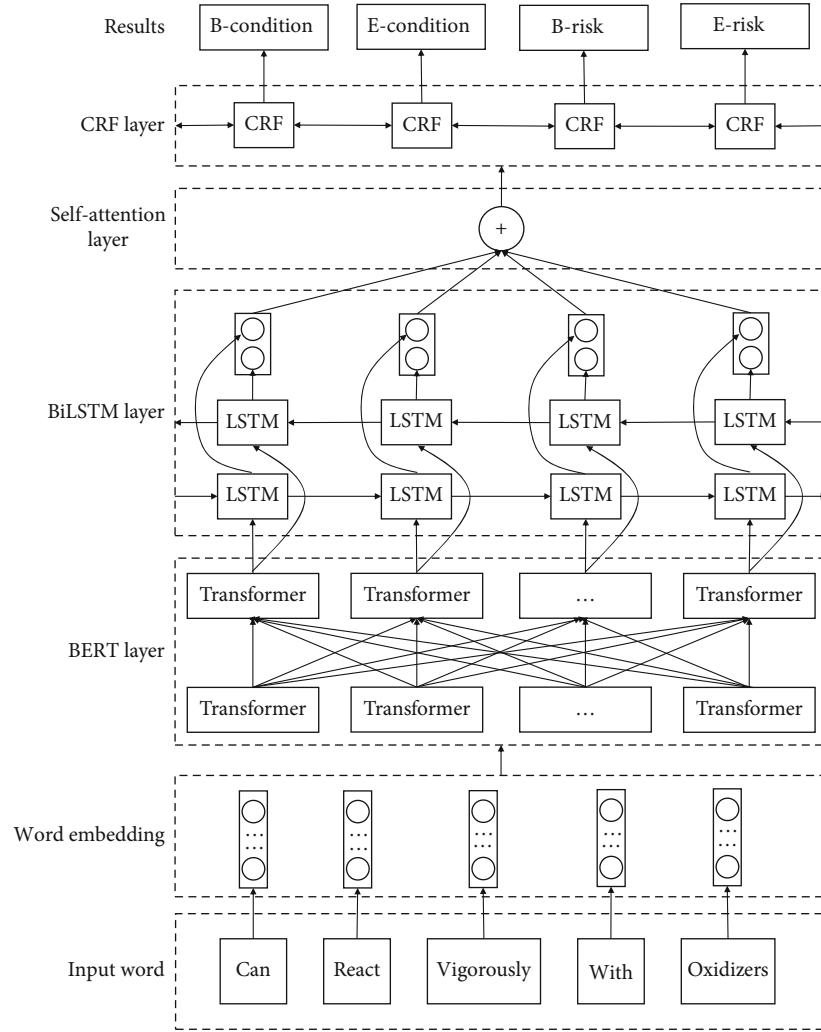


FIGURE 1: The structure of BERT-BiLSTM-Attention-CRF.

embedding models do not do anything well in word length-dependent scenarios.

The network architecture of BERT uses the transformer structure proposed by Vaswani et al. [11]. Its most notable feature is that it foregoes regular RNN and CNN in favor of using the attention mechanism to turn the word distance between any two places into 1, essentially solving the problem of widespread reliance. As shown in Figure 2, BERT's input word embedding is made up of token embedding, segment embedding, and position embedding.

When calculating self-attention, the pretrained model BERT must specify three vectors: the query vector, the key vector, and the value vector. The word embedding is multiplied by the three training weight matrices w_q , w_k , and w_v to produce these three vectors. The calculation formula of the self-attention layer is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V are the matrices of the query, key, and value vectors and d is the dimension of the input embedding. The problem of gradient disappearance may be efficiently controlled by dividing by $\sqrt{d_k}$, allowing the model's gradient to progressively drop over the training process.

Finally, using the softmax function, the score is normalized so that the output word vector may completely learn the relationship between the word and other words, enriching the semantic representation of the word.

3.2. BiLSTM Layer. RNNs contain three layers: input, hidden, and output. However, when the length of the input sequence grows longer, typical RNNs will experience gradient explosion and disappearance. LSTM is enhanced based on RNN. LSTM tackles the problems of gradient disappearance and gradient explosion by adding three gating units: forget gate, input gate, and output gate. This increases the model's convergence speed. Figure 3 depicts the construction of the LSTM.

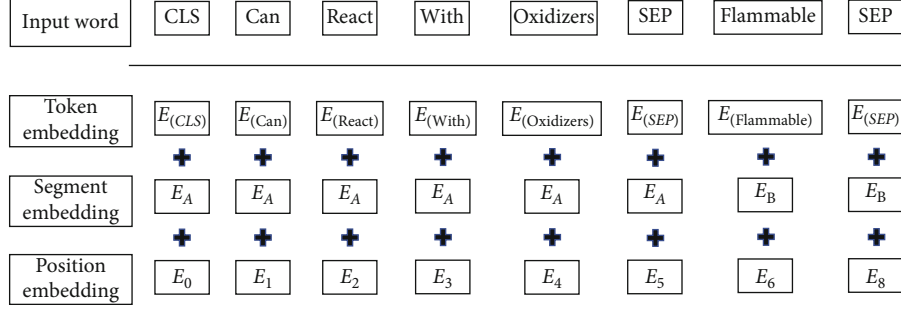


FIGURE 2: Input word embedding of BERT.

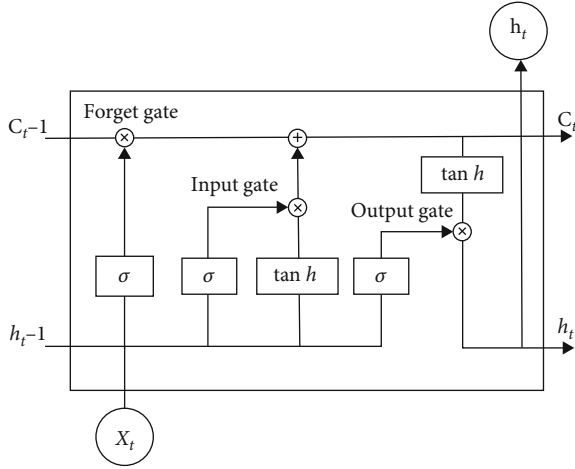


FIGURE 3: The structure of LSTM cell.

The forget gate, input gate, and output gate formulae for the LSTM unit structure are as follows:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (2)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad (3)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad (4)$$

where f_t , i_t , and o_t stand for forget gate, input gate, and output gate, respectively; W_f , W_i , and W_o represent the corresponding weight matrix, respectively; and b_f , b_i , and b_o represent the corresponding offset vector, respectively; σ represents sigmoid activation function; x_t represents the encoded input embedding at time t_n ; and h_{t-1} represents the state of the hidden layer at time t_{n-1} .

The cell state formula is as follows:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c), \quad (5)$$

where the Hadamard product is represented by \odot , \tanh stands for hyperbolic tangent activation function, W_c represents the weight matrix, and b_c represents the offset vector of the update state.

The update formula of hidden layer status is as follows:

$$h_t = o_t \odot \tanh(C_t). \quad (6)$$

Both the preceding and the following information will have an impact on the label of the current entity in the named entity recognition job. The forward LSTM can only take into account previous information in the text due to the structural properties of LSTM. As a result, this paper uses the BiLSTM model, employs bidirectional LSTM, simultaneously learns the above and below information of the text, and splices the output vector, overcoming the disadvantage of a single LSTM's incomplete semantic information and improving the accuracy of the experimental results.

3.3. Multihead Self-Attention Mechanism Layer. The labels of characters in the job of named entity identification are heavily influenced by some words in the context; therefore, the labels of the same characters might be highly different in various circumstances. The word embedding obtained by BERT cannot well express the semantics in the field of risk information about hazardous chemicals because BERT is a model trained based on ordinary text sentences. When obtaining context information, the BiLSTM network is more likely to obtain local general information, which cannot fully express the global information of the input text sequence and cannot better obtain input and local information related to the current time output. That is, it cannot fully express the importance of each character in the sentence to the current time output. Therefore, this study takes the multihead self-attention mechanism as an additional module of the BiLSTM module, which enhances the ability of the model to mine global information and sentence relevance, so that the model can be better applied in the field of risk information of hazardous chemicals.

The query vector Q , key vector K , and value vector V will employ various vector matrices for h independent linear mapping in the multihead self-attention mechanism layer and then input them into h parallel line headers to execute the self-attention operation. In this manner, each parallel line header can get the unique feature semantic information of each character in the input text sequence in a distinct presentation space. The results of the calculations on each parallel head of h heads are then merged to create a linear

TABLE 1: Description of entities in the field of hazardous chemical risk information.

Data set	Entity	Description	Example
Hazardous chemical risk information data set	SUBJECT	Risk information initiator	Nitromethane
	CONDITION	Risk information occurrence conditions	Heat
	RISK	Risk information	Dust explosion
	RESULT	Risk product or risk result	Formic acid

TABLE 2: Definition and description of hazardous chemical risk information entity category.

	Label	Meaning
1	B-SUBJECT	Subject initials
2	I-SUBJECT	Subject nonprefix and suffix
3	E-SUBJECT	Subject suffix
4	B-CONDITION	Conditional prefix
5	I-CONDITION	Conditional nonprefix and suffix
6	E-CONDITION	Conditional suffix
7	B-RISK	Risk initial
8	I-RISK	Risk nonprefix and suffix
9	E-RISK	Risk suffix
10	B-RESULT	Result initial
11	I-RESULT	Result nonprefix and suffix
12	E-RESULT	Result suffix
13	O	Others

mapping, which yields the final output. The formula for the particular function is as follows:

$$\text{head}_i = \text{attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (7)$$

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^O, \quad (8)$$

where W_i^Q , W_i^K , W_i^V , and W^O represent the weight matrix used in linear transformation, head_i represents the i th head in the multihead self-attention module, and Concat represents the splicing vector operation.

3.4. CRF Layer. BiLSTM layer and multihead self-attention mechanism layer can learn the local and global feature information between contexts and output the label of the maximum probable value of the word. However, the relationship between labels cannot be learned, resulting in the output of continuous labels that are not in conformity with logic. There are problems of disordered order of labels of the same type or wrong matching of different labels, such as connecting I-SUBJECT to B-SUBJECT. To fully learn the dependency between adjacent tags, at the end of the model, CRF is used to decode the feature information output by the multihead self-attention mechanism layer to obtain the tag sequence of the text.

In the conditional random field of a linear chain, the characteristic functions are mainly classified into two types. The first type is the state characteristic function defined on node x , which is only related to the current node; the other

TABLE 3: Composition of hazardous chemical risk information data set.

Dataset partition	Count of sentences	Count of words
Training set	3216	64.8 k
Test set	360	7.2 k

TABLE 4: Model hyperparameter settings.

Model	Hyperparameter	Value
Default	Learning rate	0.001
	Seed	1234
	Epochs	20
	Batch size	32
	Embedding size	128
+ BiLSTM	Grad norm	5.0
	Hidden size	384
	Kernel num	8
+ IDCNN	Kernel size	(2, 3, 4)
	Dilation	2
	Hidden size	768
+ BERT	Seq length	128
	Batch size	24
	Learning rate	$3e-5$
	Epoch	5
	Seed	42
+ Self-Attention	Head num	8

is the transfer characteristic function created in the node y context, which is only related to the current node and the previous node. For a given input sequence $X = x_1, x_2, \dots, x_n$, the output tag sequence $Y = y_1, y_2, \dots, y_n$ can be obtained.

The scoring function of the tag sequence can be expressed as

$$\text{score}(X, y) = \sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i), \quad (9)$$

where t_k represents the local characteristic function; s_l represents the node characteristic function; λ_k, μ_l is the weight coefficient of t_k and s_l , respectively; k stands for the number of transition characteristic functions; and l stands for the state characteristic functions.

TABLE 5: Comparison of model entity category recognition results (unit: %).

Entity	Model	Accuracy	Recall	F1
SUBJECT	IDCNN-CRF	69.23	75.00	72.00
	BiLSTM-CRF	50.00	50.00	50.00
	BERT-CRF	60.00	50.00	54.55
	BERT-BiLSTM-CRF	75.00	50.00	60.00
	BERT-BiLSTM-Self-Attention-CRF	75.00	50.00	60.00
CONDITION	IDCNN-CRF	89.94	88.82	89.38
	BiLSTM-CRF	86.87	92.47	89.58
	BERT-CRF	91.62	94.09	92.84
	BERT-BiLSTM-CRF	91.15	94.09	92.59
	BERT-BiLSTM-Self-Attention-CRF	92.67	95.16	93.90
RISK	IDCNN-CRF	92.86	90.70	91.76
	BiLSTM-CRF	95.41	96.30	95.85
	BERT-CRF	96.26	95.37	95.81
	BERT-BiLSTM-CRF	96.26	95.37	95.81
	BERT-BiLSTM-Self-Attention-CRF	95.41	96.30	95.85
RESULT	IDCNN-CRF	68.42	86.67	76.47
	BiLSTM-CRF	91.67	91.67	91.67
	BERT-CRF	95.92	97.92	96.91
	BERT-BiLSTM-CRF	97.96	100.00	98.97
	BERT-BiLSTM-Self-Attention-CRF	97.92	97.92	97.92
All	IDCNN-CRF	88.36	88.69	88.52
	BiLSTM-CRF	89.47	92.82	91.11
	BERT-CRF	93.18	94.25	93.71
	BERT-BiLSTM-CRF	93.47	94.54	94.00
	BERT-BiLSTM-Self-Attention-CRF	94.03	95.11	94.57

The scores of all feasible tag sequences are calculated for a given input sequence X . The following is the normalizing formula:

$$p(y|X) = \frac{\exp(\text{score}(X, y))}{\sum_{\tilde{y} \in Y_x} \exp(\text{score}(X, \tilde{y}))}, \quad (10)$$

where $\text{score}(X, \tilde{y})$ represents the score of the scoring function of the prediction sequence \tilde{y} , \tilde{y} represents the real annotation sequence, and Y_x represents all conceivable dimension sequences.

Finally, the Viterbi algorithm is used to get the optimum prediction tag sequence:

$$y^* = \text{argmax}(\text{score}(X, y)). \quad (11)$$

By adding a CRF layer at the end of the model, some constraints are added to the last predicted label to ensure that the predicted label is legal, so as to improve the accuracy of the predicted label.

4. Experience

4.1. Dataset and Annotation System. Resulting from various text styles and different formats of hazardous chemical risk

information, this paper takes 2828 hazardous chemicals recorded in the catalog of hazardous chemicals (2015 Edition) as the object. By crawling the material safety data sheets (MSDS) corresponding to this hazardous chemical risk information and then filtering out useless information such as pictures and repeated information through data cleaning and preprocessing, a corpus in the field of hazardous chemical risk information is constructed.

Because there are many relationships in the crawled data, the relationship is seen as a distinct entity, and the step of forming triples is combined into one step by first identifying the entity and then extracting the relationship [12]. This allows us to take full advantage of the characteristics of both entities and relationships in corpus sentences while also speeding up the construction of triples. Table 1 shows the definition of entity types in the field of hazardous chemical risk information.

The BIEO labeling approach is used to label data in this study, with B-Label representing the start of the labeled entity, I-Label representing the middle portion of the labeled entity, E-Label representing the end section of the labeled entity, and O representing unrelated information. Table 2 shows the definition of entity labels.

YEDDA [13] provides a systematic solution for text range annotation, from collaborative user annotation through

administrator assessment and analysis, as a lightweight, efficient, and complete open-source application for text span annotation. This paper removes the unnecessary parts of YEDDA and establishes an auxiliary entity annotation platform based on this experiment. Part of the data in the dataset is manually marked and divided into training set and test set according to the ratio of 9:1. Then, according to the replacement proportion of 20%, replace the similar types of entities in the training set to generate new statements and add them to the data set. Finally, supervised learning is used to mark the remaining sentences, and manual evaluation ensures that they are correctly marked. The composition of the data set is shown in Table 3.

4.2. Experimental Details. The code used in our experiment is based on the code published by Xu et al. [14], in which various models are added for comparative experiments. Table 4 shows the experimental hyperparameters of each model.

4.3. Experimental Results and Analysis. Experimental comparisons using IDCNN-CRF [2], BiLSTM-CRF [3], BERT-CRF [4], and BERT-BiLSTM-CRF [5] were undertaken to compare and validate the recognition impact of the BERT-BiLSTM-Self-Attention-CRF model on each kind of entity. Table 5 shows the outcomes of the experiment. The BERT-BiLSTM-Self-Attention-CRF model outperforms other models in identifying condition entities, risk entities and average performance. It is worth mentioning that the IDCNN-CRF model outperforms other models in identifying SUBJECT entity categories with a limited number of entities, owing to IDCNN's stronger corpus information extraction capacity than the BiLSTM model in small samples. Because the majority of the consequence entities in the corpus are long words, the BERT-BiLSTM-CRF model's result entity recognition effect is better than that of other models and even better than that of the model with a multihead self-attention mechanism layer. The attention information included in each letter in the lengthy words superimposes and impacts each other through the multi-head self-attention mechanism layer, which has an impact on the category of recognized entities. The condition entities and risk entities are mostly short words, and all BERT models with a multihead self-attention mechanism layer perform better.

5. Conclusion

Based on the hazardous chemical risk information dataset, a named entity recognition model of BERT-BiLSTM-Self-Attention-CRF is proposed. The pretrained language model BERT is introduced into the BiLSTM-CRF model to enrich the semantic features of the initial word vector. BERT is used to obtain the initialized word vector, which overcomes the problem of a lack of corpus in the field of hazardous chemicals. The multihead self-attention mechanism is utilized to capture the internal correlation between word vectors and pay more attention to the important information with high correlation. Experimental results on the hazardous chemical dataset show that BERT-BiLSTM-Self-Attention-CRF can

well identify the entities of hazardous chemicals, with an accuracy rate of 94.03%, a recall rate of 95.11%, and an F1 value of 94.57%.

At present, there are too few datasets related to hazardous chemicals, and there is no relatively complete knowledge graph of hazardous chemicals. In future research, it is necessary to further expand and improve the dataset of hazardous chemicals, extract events on this basis, and build a knowledge graph for hazardous chemicals.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no competing interest.

Acknowledgments

This work is supported by the Zhejiang Science and Technology Plan Project of China (No. 2020C03091) and the Zhejiang Provincial Central Government Guided Local Science and Technology Development Project (No. 2020ZY1010).

References

- [1] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022.
- [2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," 2018, <https://arxiv.org/abs/1810.04805>.
- [3] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," 2017, <https://arxiv.org/abs/1702.02098>.
- [4] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Computational intelligence magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [5] T. Xie, J. Yang, and L. Hui, "Chinese entity recognition based on BERT-BiLSTM-CRF model," *Computer Systems & Applications*, vol. 29, no. 7, pp. 48–55, 2020.
- [6] B. Xie, G. Shen, C. Guo, and Y. Cui, "The named entity recognition of Chinese cybersecurity using an active learning strategy," *Wireless Communications and Mobile Computing*, vol. 2021, 11 pages, 2021.
- [7] A. Klie, B. Y. Tsui, S. Mollah et al., "Increasing metadata coverage of SRA BioSample entries using deep learning-based named entity recognition," *Database*, vol. 2021, 2021.
- [8] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using BERT in the machine reading comprehension framework," *Journal of Biomedical Informatics*, vol. 118, p. 103799, 2021.
- [9] Y. Tian, X. Sun, H. Yu, Y. Li, and K. Fu, "Hierarchical self-adaptation network for multimodal named entity recognition in social media," *Neurocomputing*, vol. 439, pp. 12–21, 2021.

- [10] T. Ma, H. Zhou, Y. Tian, and N. al-Nabhan, "A novel rumor detection algorithm based on entity recognition, sentence reconfiguration, and ordinary differential equation network," *Neurocomputing*, vol. 447, pp. 224–234, 2021.
- [11] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [12] L. Ben and J. Donghong, "Joint extraction of drug entity and drug-drug interaction," *Computer Engineering and Design*, vol. 365, no. 5, pp. 1377–1381, 2017.
- [13] J. Yang, Y. Zhang, L. Li, and X. Li, "A lightweight collaborative text span annotation tool," 2017, <https://arxiv.org/abs/1711.03759>.
- [14] L. Xu, Q. Dong, Y. Liao et al., "CLUENER2020: fine-grained named entity recognition dataset and benchmark for Chinese," 2020, <https://arxiv.org/abs/2001.04351>.