frontiers | Frontiers in Chemistry

TYPE Original Research
PUBLISHED 08 February 2023
DOI 10.3389/fchem.2023.958002

OPEN ACCESS

EDITED BY
Lian Xiang Luo,
Guangdong Medical University, China

REVIEWED BY
Pengyong Li,
Xidian University, China
Weicheng Li,
University of California, San Francisco, United States

*CORRESPONDENCE
Hao Liu,
liu.hao@ouc.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

RECEIVED 31 May 2022
ACCEPTED 24 January 2023
PUBLISHED 08 February 2023

CITATION
Ma X, Yu R, Gao C, Wei Z, Xia Y, Wang X and Liu H (2023), Research on named entity recognition method of marine natural products based on attention mechanism. *Front. Chem.* 11:958002.
doi: 10.3389/fchem.2023.958002

Research on named entity recognition method of marine natural products based on attention mechanism

Xiaodong Ma[1†], Rilei Yu[1†], Chunxiao Gao[1], Zhiqiang Wei[1,2], Yimin Xia[1], Xiaowei Wang[1] and Hao Liu[1,2*]

[1]College of Computer Science and Technology, Ocean University of China, Qingdao, China, [2]Pilot National Laboratory for Marine Science and Technology, Qingdao, China

Marine natural product (MNP) entity property information is the basis of marine drug development, and this entity property information can be obtained from the original literature. However, the traditional methods require several manual annotations, the accuracy of the model is low and slow, and the problem of inconsistent lexical contexts cannot be solved well. In order to solve the aforementioned problems, this study proposes a named entity recognition method based on the attention mechanism, inflated convolutional neural network (IDCNN), and conditional random field (CRF), combining the attention mechanism that can use the lexicality of words to make attention-weighted mentions of the extracted features, the ability of the inflated convolutional neural network to parallelize operations and long- and short-term memory, and the excellent learning ability. A named entity recognition algorithm model is developed for the automatic recognition of entity information in the MNP domain literature. Experiments demonstrate that the proposed model can properly identify entity information from the unstructured chapter-level literature and outperform the control model in several metrics. In addition, we construct an unstructured text dataset related to MNPs from an open-source dataset, which can be used for the research and development of resource scarcity scenarios.

KEYWORDS

named entity recognition, marine natural products, attention mechanism, inflated convolutional neural network, conditional random field

1 Introduction

The special high-salt, high-pressure, and low-temperature living environment of the ocean has led to the formation of many marine natural products (MNPs) with novel structures and unique effects, and the drugs developed using these MNPs have unique effects on antitumor, antioxidant, and immunity enhancement (Lu et al., 2021). So far, more than 38,000 MNPs with novel structures and diverse activities have been discovered (Li et al., 2022a). However, in the early stage of drug research and development based on MNPs, it is extremely dependent on obtaining relevant research data from various data sources and numerous documents. Its application plays a key role in deepening the connection between entities and attributes. Therefore, the optimization of named entity recognition methods is the top priority of research in the early stages of marine drug development (Ghareeb et al., 2020).

The number of MNP literature is growing rapidly; for example, PubMed (Ma et al., 2022) contains more than 13,000 articles with the keyword "Marine Natural Products," and MarinLit

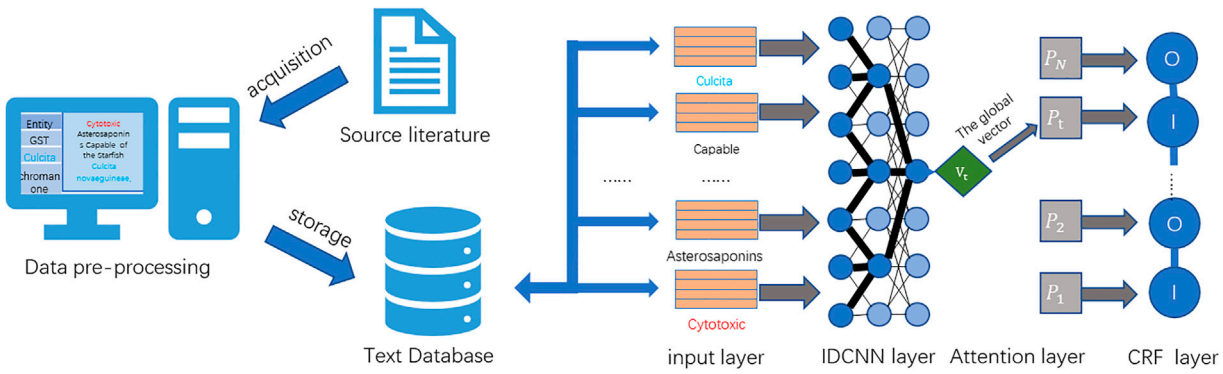

FIGURE 1
Overall technology roadmap.

(Zhang et al., 2022) contains more than 38,000 articles. However, few high-quality datasets specialize in MNP research, and there are problems such as limited data types and numbers and a lack of annotations (Lample et al., 2016). Therefore, it is particularly important to automatically obtain the attribute information on entities from the original literature and label them by themselves (Bonner et al., 2021). The application of Named Entity Recognition (NER) technology greatly improves the extraction efficiency of key information in the field of MNPs. Named entity recognition is designed to automate the identification of entity information in a domain-specific text, which can further reduce the workload of researchers in data processing and application (Cong et al., 2018). However, named entity recognition only categorizes the recognized entity and attribute information and cannot closely link entities to entities and entities to attributes. The current named entity recognition methods need to be strengthened for consistent recognition of contextually annotated entities.

Knowledge graph technology is very important for marine drug development and drug-assisted design, and the use of knowledge graph technology can greatly improve the efficiency of new drug development and reduce the cost of drug development (Sang et al., 2018). According to the survey, data in the field of marine natural products are characterized by multiple sources, heterogeneity, ambiguity, and other aspects. As a large network connecting various semantic relationships between entities and concepts, knowledge graphs can help unify and standardize entities encoded by different identifiers and fuse data from multiple heterogeneous sources. In addition, as knowledge graphs emphasize the coverage of entities, they can keep the number of points and edges huge to reach a relatively large scale in the face of the huge scale of MNP domain data. However, the current MNP knowledge graph construction mainly relies on the annotated entities or relationships and rarely involves the automated direct acquisition of entities, attributes, and attribute values from unstructured text, which is difficult to meet the requirements of building large-scale knowledge graphs (Shao et al., 2021).

Moreover, there are mainly three difficulties in the field of MNPs using named entity recognition technology to extract entities in the literature:

(1) In terms of the construction of MNP datasets, MNP data information has a wide range of sources and various data types. It is difficult to obtain relevant field information data from multiple data sources and integrate them.
(2) In terms of the textual features of MNPs, there are inconsistencies in the abbreviations, proper nouns, conjunctions, and full-text annotations in the related literature, resulting in a low recognition effect.
(3) From the perspective of the recognition accuracy of named entities in the MNP literature, the current mainstream named entity recognition methods in the field of biomedicine are inconsistent in the recognition of context-annotated entities in several MNP literature studies.

In order to solve the aforementioned problems, this study proposes an attention mechanism-based named entity recognition technology to serve the field of MNPs. First, the need for named entity recognition in the extensive literature in the field of MNPs is studied. A high-quality dataset of MNPs is constructed combined with data from MNPs and related biomedical fields. Then, various deep learning-based named entity recognition methods (Luo et al., 2017; Yadav et al., 2019; Li et al., 2021) are studied, and an ATT-IDCNN-CRF method is improved and proposed. The method has good performance in terms of training efficiency and recognition effect on the dataset proposed in this study and basically solves the difficulty of named entity recognition in the field of MNPs. The overall technology roadmap for this study is shown in Figure 1.

2 Related work

MNP databases are an important reference for marine drug discovery and development, and the number of databases is gradually increasing. MarinLit and the Marine Natural Products Dictionary are currently the most exhaustive and complete MNP databases. However, the need for subscriptions and the limitation to payment reduce the breadth of academic research. Marin Chem 3D (July 2018) is the world's first three-dimensional structure database of MNPs. Although there are not much activity data, it contains more than 30,000 three-dimensional structures of MNPs. CMNPD (Lyu et al., 2021) is also a very comprehensive database of more than 31,000 chemical entities with a wide range of physical, chemical, and

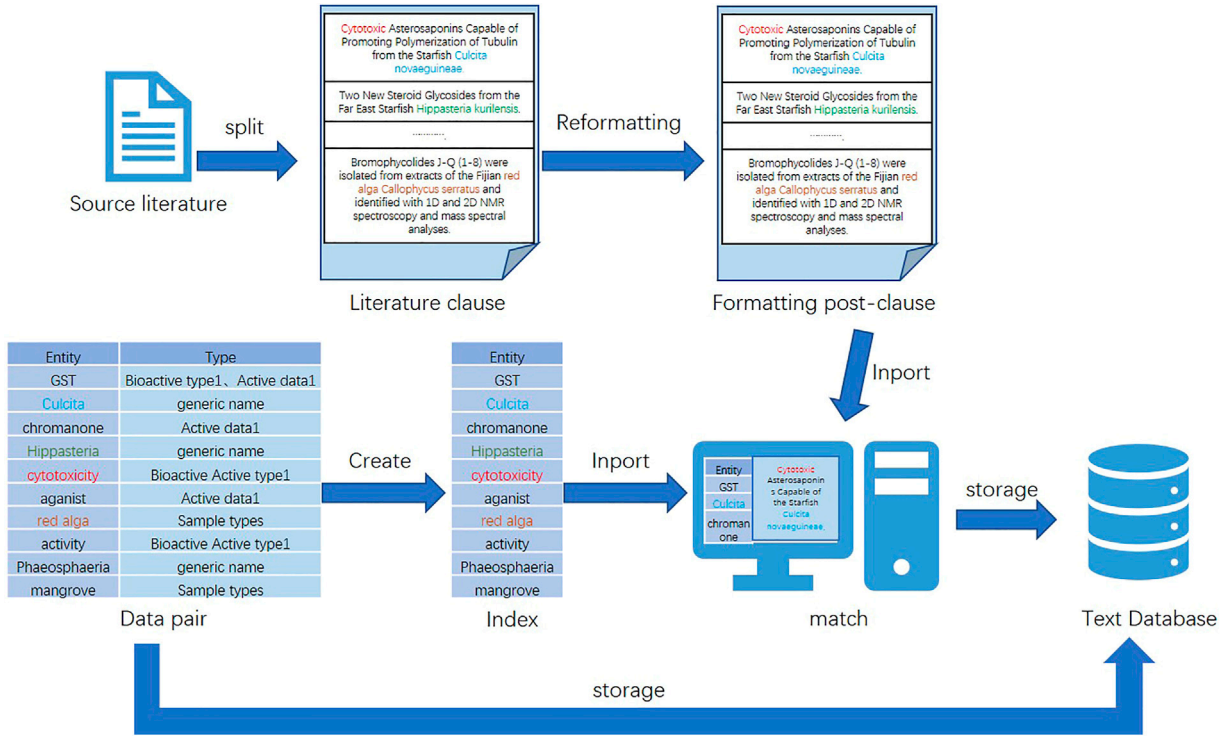

FIGURE 2
Dataset preparation process.

pharmacokinetic properties; standardized bioactivity data; systematic classification; and geographical distribution of source organisms and detailed literature citations, which were constructed from the initial manual acquisition of information on entity properties in the literature by experts to the later application of named entity identification methods to automate the identification of desired entities from the original literature. Although the most widely used convolution neural network (CNN) (Aslan et al., 2021) has obvious computational advantages in the named entity recognition of MNP documents, the traditional CNN can only obtain a small part of the input text information after convolution. In order to obtain contextual information, more convolutional layers need to be added, resulting in deeper networks, more parameters, and being prone to overfitting (Fang et al., 2020). Inflated convolutional neural network (IDCNN) (Li et al., 2022b) adds convolution holes to CNN, which enables IDCNN to control its sliding window to omit inputs of a specific length range, which can reduce the number of convolutional layers to better capture sentence context information and greatly improve the efficiency of parallel computing (Wang and Xu, 2017). However, its recognition accuracy for long-named entities is not high. Long short-term memory networks (LSTMs) (Ma et al., 2021) are a special type of recurrent neural networks (RNNs) that can learn long-range dependencies. Currently, the BiLSTM-CRF model based on bidirectional LSTM (BiLSTM) combined with CRF has become the most mainstream model in deep learning-based NER methods (Wu et al., 2019). In terms of features, the model inherits the advantages of deep learning methods. It can achieve good results using word and character vectors without feature engineering. If there are high-quality dictionary features, it can be further improved (Wang et al., 2018). The accuracy of the named entity recognition model will directly determine the success of knowledge graph construction and entity activity relationship prediction (Zeng et al., 2017). Knowledge graph construction is also a key step in database establishment. The construction of early knowledge graphs was manually constructed by experts in related fields, such as WordNet (Miller, 1995), Cyc (Lenat and Douglas, 1995), and OpenCyc (Färber et al., 2018), but it is extremely labor-intensive. With the establishment of the World Wide Web, it has entered the era of a semantic network, in which DBpedia (Auer et al., 2007) and Yago (Rebele et al., 2016) are the representatives that combine entities and relationships into a semantic network. However, it still needs to be built manually. The era of knowledge graphs has emerged to effectively search and analyze knowledge and apply entities in other aspects. Typical uses of knowledge graphs for MNPs include CMNPD (Lyu et al., 2021), MC3D, and Marine Chinese Medicine Knowledge Graph (Liu and Li, 2021). However, the biological attribute information in the MNP knowledge graph system remains lacking and needs to be supplemented.

This study aims to solve the construction of datasets and the optimization of named entity recognition models during the construction of the MNP knowledge graph system. We constructed an unstructured text dataset related to MNPs from open-source datasets for research and development in resource-scarce scenarios. In terms of method, a named entity recognition method based on the attention mechanism, inflated convolutional neural network (IDCNN), and CRF is proposed, which can automatically identify entity attribute information in the MNP literature using attention. The mechanism to obtain full-text-level context information improves the situation of inconsistent recognition results of the same word.

Experiments demonstrate that the proposed model can properly identify entity information from the unstructured chapter-level literature and outperform the control model on multiple metrics.

3 Construction of a marine natural product dataset

Few datasets specialize in MNP domain research, and the related data sources have high charges, imperfect data, and low retrieval efficiency. A diverse, data-rich, and strong correlation of the MNP domain dataset is created in this study to verify the effectiveness of the proposed method.

3.1 Dataset acquisition

To provide entity-relationship dependence of datasets in MNPs and related biomedical fields, we selected more than 30 existing public databases in the field of marine biomedicine for our study and finally selected the literature searched in PubMed with "Marine natural product" as the keyword. The abstracts in PubMed and some abstracts in MarinLit were selected. In addition, this study also includes some annotated data provided by the laboratory team. Therefore, in addition to investigating these databases in the field of MNPs and related pharmaceuticals, we need to use specific methods to obtain data from different data sources and organize and integrate these data to build a literature dataset in the field of MNPs.

3.2 Data preprocessing

Due to the diverse structure of the acquired data, it cannot be directly used for dataset construction and named entity recognition experiments. Converting these data to the same form is an important basis for improving the accuracy of subsequent work (Davis et al., 2011). Moreover, the experimental process of named entity recognition in this study is relatively complex, and the processing time is long. Therefore, this study preprocesses the literature used for entity extraction. In order to reduce the interference factors to the experiment and improve the accuracy and operation efficiency of the experimental results, this study excludes irrelevant sentences from the data and only retains sentences with valid entity results in the data. The specific process is as follows: first, this study splits the sentences of the literature, then performs simple formatting processing on them (e.g., adding breaks and punctuation and removing extra spaces), and omits sentences that do not have valid entities and are inappropriate in length. After that, a data pair of <entity, type > is created for entity information, and an index is built on the data pair. Finally, the indexed data pairs are retrieved and paired with the preprocessed sentences.

3.3 Data storage

After the data preprocessing is completed, the data should be stored. Generally, the dataset can be stored in a text or data table form. The database of the data table type includes a relational database. Text databases can be used to store data in any text form. The latter are mainly relational databases, type tables, and triple tables. The storage of these three types of table structures has the advantage of being easy to understand and operate. However, once the data scale becomes larger or the operation becomes more complex, the advantages of such structures will disappear. Instead, it will be characterized by high overhead, low practicability, and low efficiency. The advantage of using a text database is as follows: when faced with large-scale data, it can properly combine text algorithms to improve the efficiency of complex queries and strong information description capabilities. Therefore, in the face of the large volume of data in this study, out of consideration of performance, stability, compatibility, and other characteristics, this study chooses the text database as the storage database of the dataset. A diagram of the dataset preparation process for this study is shown in Figure 2.

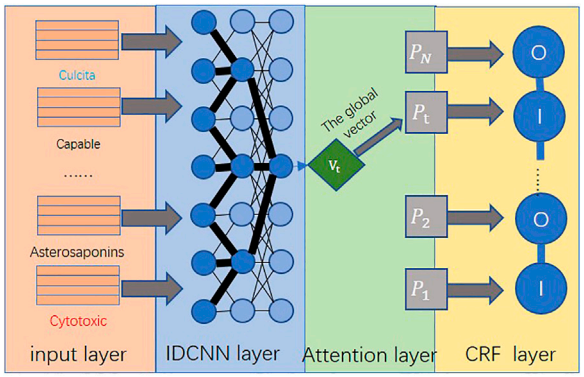

FIGURE 3
Improved ATT-IDCNN-CRF model architecture.

4 Improved named entity recognition method

4.1 IDCNN-CRF architecture

According to the actual needs, this study selects the current high-efficiency dilated convolution model (IDCNN-CRF). IDCNN increases the field of view of the convolution exponentially, but the parameters used increase linearly, which can fully utilize the power of GPU parallel computing, so compared with other models, the use of IDCNN-CRF can greatly speed up the training speed.

Similar to most traditional machine learning NER methods, the IDCNN-CRF method is a sentence-level NER method (Sagar et al., 2013). A common problem at the sentence level is that the contextually labeled entities may be inconsistent. In order to address this issue, previous work often employs a rule-based post-processing approach to enforce label consistency (Kim et al., 2019). However, if the label is misidentified in the front, it will greatly increase the misjudgment of subsequent labels (Mendez et al., 2019). In order to avoid the errors caused by the aforementioned methods, this study introduces the attention mechanism on the basis of the original IDCNN-CRF model. The attention mechanism can use the part-of-speech of words to weigh the extracted features to improve the consistency of contextual annotation entities.

TABLE 1 Complete entity types and some examples.

Entity type	Sample
Formula	$C_{60}H_{94}O_{34}S_2Na_2$ and $C_{35}H_{42}O_{11}$
Sample source	The coast of Key Largo, off Ulleung Island, Korea
Species	*Theonella* and *moluccensis*
Generic name	*Siliquariaspongia* sp. and *Xylocarpus*
Active data	MIC80, <50.0
Periodical information	Molecules 2009, 14, 414–422, J. Nat. Prod. 2009, 72, 1657–1662
Compound name	Lepirudin and S-sulfocysteine
Relative molecular mass	201.221 and 156.850
Name-in-notebook	1987–2, 1987–3
Smiles	c1cc (c2c (c1)C (=O)c1c (C2 = O)cc2c (c1O)cc (cc2O)C)O
Molecular weight	684.4210 and 320.3040
Name	Prosurugatoxin and surugatoxin
Number of heavy atoms	9, 24
Number of rotatable bonds	3
donorHB	9, 1
PSA	112.0280
AlogP	−2.9260 and 3.4010
accptHB	18, 5
Species and origin	Digestive gland of *B. japonica*
Action	Prosurugatoxin evoked mydriasis in mice at a minimum effective intraperitoneal dose of 15 ng/g body weight and inhibited the contractile response of isolated guinea pig ileum induced by 3 × 10-5 g/ml of nicotine at a concentration of 5 × 10-9 g/ml
PercentHumanOralAbsorption	0.0000 and 77.0340
PMDCK (nm/s)	0.4750 and 47.3340
Pcaco (nm/s)	0.6670 and 114.0480
logS (S in mol/L)	−3.1160 and −4.1410
logHERG (IC_{50})	−4.1280 and −5.0440
logBB	−3.7450 and −1.4770
logKp	−8.5020 and −4.1240
logKhsa	−0.8550 and 0.2310

4.2 Introduction of the attention mechanism

As the current point in the MNP literature has a low correlation with some long-distance information, even if the receptive field area of the dilated convolution is increased to obtain long-distance information, the data cannot have a high degree of accuracy consistency. In order to improve the accuracy, the named entity recognition task in this study needs to obtain full-text-level content information as much as possible. Therefore, this study uses the attention mechanism to improve the IDCNN-CRF model. First, in order to achieve the purpose of weight summation, the model will change the word vector and word vector in the way of combination. Then, in order to achieve the purpose of learning the attention weight, a hidden layer with two layers is used for combined learning. In this way, the model can use both word vector and character vector information to non-statically obtain text-level unstructured context-dependent relevant information (Luo et al., 2018).

In the process of entity extraction, the sentence in the document is first converted into a sequence of word vectors, and the sequence of word vectors is used to generate character vectors, and then, a matrix of character vectors is constructed. To obtain character-level features for each word, IDCNN applies convolution and pooling to a matrix of character vectors. By combining the word vector and character vector of each word, the input sequence of the IDCNN layer is formed, which is finally input into the network.

TABLE 2 Test input and output sample table.

Input	Output		
	Entity	Type	Start and end location
Cytotoxic asterosaponins capable of promoting polymerization of tubulin from the starfish ***Culcita novaeguineae***	Cytotoxic	Bioactivity	0 and 8
	Culcita	Generic name	90 and 96
	novaeguineae	Specific name	98 and 109
Two new steroid glycosides from the far east starfish ***Hippasteria kurilensis***	*Hippasteria*	Generic name	54 and 64
	kurilensis	Specific name	66 and 75
Bromophycolides J-Q (1–8) were isolated from extracts of the Fijian **red alga *Callophycus serratus*** and identified with 1D and 2D NMR spectroscopy and mass spectral analyses	Red alga	Sample types	68 and 75
	Callophycus serratus	Generic name	77 and 96
As part of our search for bioactive substances from marine organisms systematically and assessing the chemical and biological diversities of seaweeds distributed along the Chinese coast, the **red alga** *Laurencia similis* was collected from **Sanya Bay, Hainan province**	Red alga	Sample types	184 and 191
	Laurencia similis	Generic name	193 and 209
	Sanya Bay, Hainan province	Sample source	230 and 255
Polysiphonia urceolata was collected **at the coast of Yantai, China**, in May 2008, and identified by Prof. Xiao Fan of the Institute of Oceanology, Chinese Academy of Sciences	*Polysiphonia urceolata*	Generic name	0 and 21
	At the coast of Yantai, China	Sample source	37 and 65

TABLE 3 Comparison of experimental results of recognition models.

Model	Precision	Recall	F	Speed (ep)/s
IDCNN-CRF	88.44	86.39	87.41	76
BiLSTM-CRF	90.57	89.35	89.96	253
LSTM-CRF	89.01	88.07	88.53	187
Att-IDCNN-CRF	92.18	90.71	91.44	98

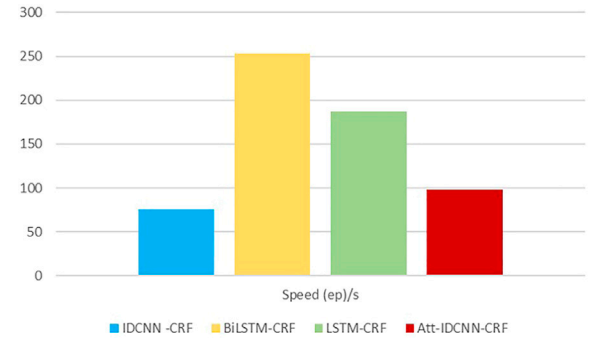

FIGURE 4
Chart of the speed metric.

After analysis, the input file D = $(X_1, X_2, \ldots, X_m)$ consists of m sentences; each sentence is represented as X = $(X_1, X_2, \ldots, X_n)$, where n is the number of words in the sentence. Like the regular IDCNN-CRF model, the embedding vector transformed using the word2vec tool is first provided as the input to the IDCNN layer. Then, in addition to the original IDCNN layer, an additional layer is added—the attention layer. The newly added layer acts as an attention mechanism in the constructed model, using the Attention layer in the document to explore the similarity between tags. To this end, the attention similarity matrix S is specially introduced here, and the attention similarity matrix S is used to calculate the similarity between the tags we are concerned about and each tag in the document. The similarity is expressed by the attention degree. Among them, in the attention matrix S, we can obtain the weight required by the attention matrix S by comparing the representation of the attention word i and the jth word in the document as follows:

$$s_{i,j} = \frac{\exp\left(f\left(x_i, x_j\right)\right)}{\sum_m^i \exp\left(f\left(x_i, x_m\right)\right)}. \tag{4–1}$$

$f(x_i, x_j)$ is the function of the outputting probability value, which acts as an alignment function. To facilitate data processing and speed up computation, we use tanh to compute the value of $f(x_i, x_j)$:

$$f\left(x_i, x_j\right) = \tanh\left(M_s\left[x_i; x_j\right]\right). \tag{4–2}$$

In the aforementioned score function calculation formula, M_s is called the weight transformation matrix, equivalent to a parameter in the model. Then, the output value q_j of each IDCNN is weighted and summed, and the weighted sum is assigned to the vector v_i acting on the whole document:

$$v_i = \sum_{j=1}^{N} s_{i,j} q_j. \tag{4–3}$$

The following step is to generate the output of the attention layer. The specific steps are as follows: first, concatenate the global vector v_i and the output value q_j of IDCNN, then feed the concatenated value to the tanh function, and finally generate the attention layer output:

$$p_i = \tanh\left(M_v\left[v_i; q_j\right]\right). \tag{4–4}$$

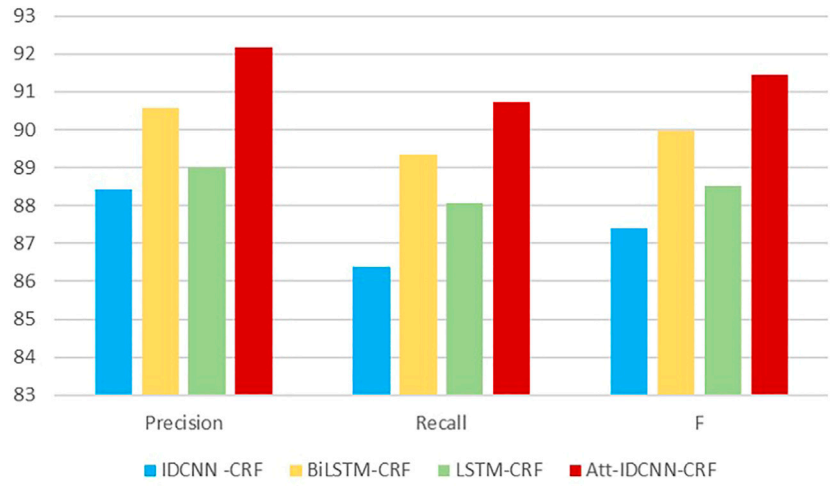

FIGURE 5
Chart of precision, recall, and F-statistic metrics.

Then, predict the activation function score for the word. This score is predicted by the tanh function value obtained in the previous step:

$$e_i = \tanh\ (S; p_i). \tag{4–5}$$

In this model, we first perform two regular 1D convolution operations on the output of the word embedding layer and then perform a dilated convolution operation to replace the max pooling layer in a standard convolutional neural network, and finally put the attention mechanism layer. The tanh layers are output as two fully connected layers.

At the end of the MNP entity extraction work, we use the CRF layer in the IDCNN-CRF model to select the best path, as shown in Eqs. 4–6, where R is the label information transfer matrix of this model and O is used as this model. The data score result matrix of the final calculated score result will be calculated by the comprehensive calculation of the input document D and y in all label paths, and an optimal label transfer path is decoded:

$$L(D, y) = \sum_m \sum_{i=1}^{n} \left(R_{y_i-1,y_i} + O_{i,y_i}\right). \tag{4–6}$$

4.3 ATT-IDCNN-CRF model

The ATT-IDCNN-CRF model constructed in this study can fully guarantee the contextual information on the text at the same time and make the model parameters not too much to cause overfitting. Therefore, the model can improve the training speed while ensuring the accuracy of text feature extraction. ATT-IDCNN-CRF is mainly composed of the attention, IDCNN, and CRF layers. The architecture of the model in this study is shown in Figure 3.

5 Result

5.1 Dataset

The training sets of the experiments in this study are mainly the MNP text dataset and the corpus CHEMDNER of BioCreative IV in the biological field (Chollet, 2017). The test set is the MNP text dataset proposed in this study. Among them, the data entity types comprise dozens of types, including unrelated samples, such as compounds, targets, small drug molecules, diseases, proteins, documents, and drugs, as well as the attributes of the aforementioned entities (Karmakar et al., 2021). The literature is divided into 37,926 training datasets and 7,586 testing datasets. Table 1 shows information about complete entity types and examples.

5.2 Evaluation criteria

Equations (5–1), (5–2), and (5–3) give the calculation method of the model evaluation index and the meaning of the corresponding parameters: T_p represents the real sample, F_p represents the false positive sample, T_n represents the true negative sample, and F_n represents a false negative sample. We define the samples whose entity types and locations in the test samples are the same as the model output judgment results as real samples; the samples whose output result is that the entity can be recognized, but the category or boundary judgment is wrong as false positive samples; the samples with no entity information in the prediction results of the entity extraction model and the test samples with no entity information as true negative samples; and the samples that do not contain entity information in the prediction results through the model output but should have entity information as false negative samples.

Select precision, recall, F value, and speed as the validation evaluation metrics for this experiment. Speed is an indicator proposed in this study to measure the running time of model training. Due to different datasets and hardware environments, the running time will be different. Therefore, in the following comparison experiments, the epoch of the total training should be controlled to be the same, and the speed value is the average training time (in seconds) required for each epoch:

$$\text{Precision} = \frac{T_p}{T_p + F_p}, \tag{5 – 1}$$

$$Recall = \frac{T_p}{T_p + F_n}, \quad (5-2)$$

$$F = \frac{2*Precision*Recall}{Precision + Recall}. \quad (5-3)$$

5.3 Analysis of results

The input for this experimental test is a chapter-based document sentence, and the output is the entity, the entity type, the start position of the entity, and the end position of the entity. The input and output samples are shown in Table 2.

The experimental results of this experiment are shown in Table 3. After analysis and comparison, the following conclusions can be drawn: compared with LSTM, BiLSTM-CRF, and other models applied to the named entity extraction, the improved attention-IDCNN-CRF on the original basis of the speed value of the model in the identification task of MNP source information is much higher than that of models LSTM-CRF and BiLSTM-CRF on average, as shown in Figure 4, and the recognition accuracy is significantly improved. Comprehensive experimental comparison results basically realized the main research objective of this study.

Compared with the IDCNN-CRF model, the ATT-IDCNN-CRF model has a slight decrease in recognition speed. However, the attention mechanism can pay more attention to the weight of the chapter level than the traditional word/word level vector, and the F value is improved in the experiments of this study. Figure 5 shows that the IDCNN-CRF model with the attention mechanism has a better semantic expression ability.

6 Conclusion

This is a systematic study of named entity recognition methods for the MNP literature. First, it constructs a dataset of unstructured text in the field of MNPs. Second, an attention-based IDCNN-CRF named entity recognition model is improved and trained. By comparing multiple indicators, the advantages of the model are verified on the MNP dataset. The unique compound data information on MNPs is of great significance to related medical research, and it is very important to automatically extract relevant information. In the future, we will conduct more in-depth research in the field of named entity recognition of MNPs so that the training will be faster and the recognition effect will be more accurate.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Author contributions

XM re-optimized the design of the model, conducted experiments, and edited the manuscript. RY suggested model optimization design and data collection, and guided some experiments. CG participated in the collection of experimental data and the writing of some codes. YX and XW participated in the implementation of the model experiment. ZW proposed revisions to the paper. HL optimized the model, designed the overall framework of the system, and revised the paper.

Funding

This work was supported by the National Key Research and Development Program of China (2021YFF0704000) and the National Natural Science Foundation of China (NSFC) (nos. 32171267 and 82122064).

Acknowledgments

The authors would like to thank Jie Liu, Hao Lu, and Yangyang Li for their suggestions and discussions. They also thank the Ocean University of China for their support of the experiments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aslan, M. F., Fahri Unlersen, M., Sabanci, K., and Durdu, A. (2021). CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection. *Applied Soft Computing*. 98, 106912. doi:10.1016/j.asoc.2020.106912

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., et al. (2007). "DBpedia: A nucleus for a Web of open data," in *The semantic Web*. K. Aberer, K.-S. Choi, N. Noy, A. Dean, K.-I. Lee, L. Nixon, et al. Editors (Berlin, Heidelberg: Springer), 722–735. Lecture Notes in Computer Science.

Bonner, S., Barrett, I. P., Cheng, Y., Swiers, R., Engkvist, O., Bender, A., et al. (2021). A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Available at: https://arxiv.org/abs/2102.10062.

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in In 30th IEEE Conference on Computer Vision and Pattern Recognition CVPR 2017, Honolulu, United states, July 21, 2017 - July 26, 2017. 2017-January:1800–1807. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. (Institute of Electrical and Electronics Engineers Inc. doi:10.1109/CVPR.2017.195

Cong, Q., Feng, Z., Li, F., Zhang, L., Rao, G., and Cui, T. (2018). "Constructing biomedical knowledge graph based on SemMedDB and linked open data," in In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, December 2018, 1628–1631. doi:10.1109/BIBM.2018.8621568

DavisJohn, G. D., and Hannah Rachel Vasanthi, A. (2011). Seaweed metabolite database (swmd): A database of natural compounds from marine algae. *Bioinformatics* 5 (8), 361–364. doi:10.6026/97320630005361

Fang, Y., Gao, J., Liu, Z., and Huang, C. (2020). Detecting cyber threat event from twitter using IDCNN and BiLSTM. *Applied Sciences*. 10 (17), 5922. doi:10.3390/app10175922

Färber, M., Bartscherer, F., Menne, C., and Rettinger, A. (2018). Linked data quality of DBpedia, freebase, OpenCyc, wikidata, and YAGO. *Semantic Web* 9 (1), 77–129. doi:10.3233/SW-170275

Ghareeb, M. A., Tammam, M. A., Amr El-Demerdashand Atanasov, A. G. (2020). Insights about clinically approved and preclinically investigated marine natural products. *Current Research in Microbiology and Biotechnology*. 2, 88–102. doi:10.1016/j.crbiot.2020.09.001

Karmakar, P., Teng, S. W., and Lu, G. (2021). Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. arXiv: 2102.07259. arXiv Available at: https://arxiv.org/abs/2102.07259.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., Jia, H., He, S., et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* 47 (D1), D1102–D1109. doi:10.1093/nar/gky1033

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). "Neural architectures for named entity recognition," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, California, United States, June 2016. doi:10.48550/ARXIV.1603.01360

Lenat,and Douglas, B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*. 38 (11), 33–38. doi:10.1145/219717.219745

Li, Y., Li, X., Ma, P., and Ma, J. (2021). "Overview: The databases of chemical components of traditional Chinese medicine," in In 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, May 2021, 32–37. doi:10.1109/ICAIBD51990.2021.9459017

Li, J., Sun, A., Han, J., and Li, C. (2022a). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*. 34 (1), 50–70. doi:10.1109/TKDE.2020.2981314

Li, M., Zhou, G., and Lu, C. (2022b). Peach surface defect identification of complex background based on IDCNN and GWOABC-KM. *Multimedia Tools and Applications*. 81 (12), 16309–16334. doi:10.1007/s11042-022-12563-2

Liu, L., and Li, X. (2021). "Research and construction of marine Chinese medicine formulas knowledge graph," in In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, United States, December 2021, 3853–3855. doi:10.1109/BIBM52615.2021.9669655

Lu, W.-Y., Li, H.-J., Li, Q.-Y., and Wu, Y.-C. (2021). Application of marine natural products in drug research. *Bioorganic and Medicinal Chemistry*. 35 (April), 116058. doi:10.1016/j.bmc.2021.116058

Luo, Y., Lu, Y., Wang, L., and Cheng, H. (2017). "Efficient CNN-CRF network for retinal image segmentation," in *Cognitive systems and signal processing*. Editors F. Sun, H. Liu, and D. Hu (Gateway East, Singapore: Springer). 157–65Communications in Computer and Information Science.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., et al. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* 34 (8), 1381–1388. doi:10.1093/bioinformatics/btx761

Lyu, C., Chen, T., Qiang, B., Liu, N., Wang, H., Zhang, L., et al. (2021). Cmnpd: A comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Research*. 49 (D1), D509–D515. doi:10.1093/nar/gkaa763

Ma, R., Zheng, X., Wang, P., Liu, H., and Zhang, C. (2021). The prediction and analysis of COVID-19 epidemic trend by combining LSTM and markov method. *Scientific Reports*. 11 (1), 17421. doi:10.1038/s41598-021-97037-5

Ma, J., Wu, X., and Huang, L. (2022). The use of artificial intelligence in literature search and selection of the PubMed database. *Rahman Ali Science Program*. 2022, 1–9. doi:10.1155/2022/8855307

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*. 47 (D1), D930–D940. doi:10.1093/nar/gky1075

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*. 38 (11), 39–41. doi:10.1145/219717.219748

Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., and Weikum, G. (2016). "Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames," in *The semantic Web – ISWC 2016*. Editors P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, et al. (Gateway East, Singapore: Springer International Publishing). 177–85. Lecture Notes in Computer Science.

Sagar, S., Kaur, M., Radovanovic, A., and Bajic, V. B. (2013). Dragon exploration system on marine sponge compounds interactions. *Journal of Cheminformatics* 5 (1), 11. doi:10.1186/1758-2946-5-11

Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H., and Wang, J. (2018). SemaTyP: A knowledge graph based literature mining method for drug discovery. *BMC Bioinformatics*. 19 (1), 193. doi:10.1186/s12859-018-2167-5

Shao, B., Li, X., and Bian, G. (2021). A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. *Expert Systems with Applications*. 165, 113764. doi:10.1016/j.eswa.2020.113764

Wang, C., and Xu, B. (2017). Convolutional neural network with word embeddings for Chinese word segmentation. arXiv:1711.04411. arXiv Available at: https://arxiv.org/abs/1711.04411.

Wang, X., Yang, R., Lu, Y., and Wu, Q. (2018). "Military named entity recognition method based on deep learning," in In 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, November 2018, 479–483. doi:10.1109/CCIS.2018.8691316

Wu, G., Tang, G., Wang, Z., Zhang, Z., and Wang, Z. (2019). An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. *IEEE Access* 7, 113942–113949. doi:10.1109/ACCESS.2019.2935223

Yadav, V., and Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. arXiv:1910.11470. arXiv Available at: https://arxiv.org/abs/1910.11470.

Zeng, D., Sun, C., Lin, L., and Liu, B. (2017). LSTM-CRF for drug-named entity recognition. *Entropy* 19 (6), 283. doi:10.3390/e19060283

Zhang, S., Song, W., Nothias, L.-F., Couvillion, S. P., Webster, N., and Torsten, T. (2022). Comparative metabolomic analysis reveals shared and unique chemical interactions in sponge holobionts. *Microbiome* 10 (1), 22. doi:10.1186/s40168-021-01220-9