Article

Joint Entity-Relation Extraction for Knowledge Graph Construction in Marine Ranching Equipment

Du Chen , Zhiwu GAO , Sirui Li , Xuruixue Guo , Yaqi Wu , Haiyu Zhang [*] , Delin Zhang [*]

Article

Joint Entity-Relation Extraction for Knowledge Graph Construction in Marine Ranching Equipment

Du Chen [1], Zhiwu Gao [1], Sirui Li [1] and Xuruixue Guo [1],Yaqi Wu [1],Haiyu Zhang [1],*,Delin Zhang [2],*

[1]   Yantai Institute, China Agricultural University, Yantai 264670, China
[2]   College of Water Resources and Civil Engineering,China Agricultural University, Beijing 100083, China
*   Correspondence: zhanghaiyu@cau.edu.cn (H.Z);zdl188@cau. edu. cn(D.Z)

Abstract: The construction of marine ranching is a crucial component of China's Blue Granary strategy, yet the fragmented knowledge system in marine ranching equipment impedes intelligent management and operational efficiency. This study proposes a novel knowledge graph (KG) framework tailored for marine ranching equipment, integrating hybrid ontology design, joint entity-relation extraction, and graph-based knowledge storage: (1) The limitations in existing KG are obtained through targeted questionnaires for diverse users and employees; (2)A domain ontology was constructed through a combined top-down and bottom-up approach, defining seven core concepts and eight semantic relationships; (3) Semi-structured data from enterprises and standards, combined with unstructured data from literature were systematically collected, cleaned via Scrapy and regular expression, and standardized into JSON format, forming a domain-specific corpus of 1,456 annotated sentences; (4)A novel BERT-BiGRU-CRF model was developed, leveraging contextual embeddings from BERT, parameter-efficient sequence modeling via BiGRU, and label dependency optimization using CRF. The TE+SE+R_i+BMESO tagging strategy was introduced to address multi-relation extraction challenges by linking theme entities to secondary entities; (4)The Neo4j-based knowledge graph encapsulated 2,153 nodes and 3,872 edges, enabling scalable visualization and dynamic updates. Experimental results demonstrated superior performance over BERT-BiLSTM-CRF, achieving 86.58% precision, 77.82% recall, and 81.97% F1-score(1.94% improvement, $p < 0.05$) . This study not only pioneers the first structured KG framework for marine ranching equipment but also offers a transferable methodology for vertical domain knowledge extraction.

Keywords: Marine ranching equipment; Knowledge graph; Hybrid ontology; BERT-BiGRU-CRF; Joint entity-relation extraction

1. Introduction

Marine ranching, a cornerstone of China's Blue Granary strategy, has emerged as a transformative approach to modernize marine fisheries, enhance aquaculture productivity, and promote ecological sustainability [1]. The rapid development of related equipment, such as intelligent feeding systems, deep-sea cages, and multi-functional platform, have significantly improved operational efficiency [2,3].However, the exponential growth of domain-specific knowledge remains fragmented, with critical information dispersed across heterogeneous sources including enterprise records, experts experiences, academic literature, and technical standards.This fragmentation impedes intelligent decision-making, real-time monitoring, and knowledge sharing, thereby limiting the full potential of marine ranching industrialization.

Knowledge graph (KG) is a state-of-the-art semantic network paradigm, it employs graph structures to visualize relationships between entities, demonstrating advantages in intuitiveness, efficiency, and scalability[4]. The concept of KG was first proposed by Google in 2012 and applied in the search engine field[5].It has demonstrated remarkable success in integrating domain-specific

knowledge, and can provide robust support for intelligent applications such as question-answering systems and decision-making analytics. Recent advancements in knowledge graph (KG) applications demonstrate their versatility in manufacturing, Ren et al. [6] automated OPC UA information modeling via KG to unify heterogeneous equipment data, while Gu et al. [7] integrated geometric and assembly process data through a KG-based semantic model (KG-ASM). For design optimization, Huet et al. [8] proposed a KG-driven design rule recommendation system, and Qin et al. [9] developed a KG-embedded tool for mechanical component design assistance. In quality management, Zhou et al. [10] utilized manufacturing data to construct KG for defect root-cause analysis, and Hao. [11] established a production-process-supported KG for anomaly detection.

Despite these advancements, existing research predominantly focuses on structured data from product design or assembly processes, neglecting the unique challenges of marine equipment domains where unstructured text dominates and entities exhibit complex interdependencies. Traditional extraction methods suffer from error propagation and inefficiency in handling such scenarios, while deep learning models like BERT-BiLSTM-CRF face limitations in parameter efficiency and contextual dependency modeling. By designing targeted questionnaires for diverse users and employees, several limitations in existing KG have been identified as follows: （1）Limited data volume and limited knowledge scope.（2）Ambiguity in the structure of the knowledge framework（3）Low efficiency and accuracy in knowledge extraction.（4）Difficult updates and maintenance.Consequently, there is an urgent need for a specialize KG framework tailored to marine ranching equipment.

This study proposes a novel KG framework to bridge the gap between unstructured marine equipment data and knowledge. The main contributions of this study include:

(1)Hybrid Ontology Design: A combined top-down and bottom-up approach constructs a domain ontology, defining 7 core concepts and 8 semantic relationships .

(2) Joint Extraction Model: A BERT-BiGRU-CRF model that integrating BERT's contextual embeddings, BiGRU's parameter-efficient sequence modeling and CRF's global label optimization was developed. A novel TE+SE+R_i+BMESO tagging strategy resolves multi-relation extraction challenges.

(3)Dynamic Knowledge Storage: The extracted triples are stored in Neo4j, enabling scalable visualization and real-time updates via Cypher queries.

This work pioneers the first KG framework for marine ranching equipment, offering a transferable solution for vertical domains. By transforming fragmented data into structured knowledge, our framework supports intelligent applications including equipment fault diagnosis, maintenance planning, and policy formulation.

The remainder of this paper is structured as follows: Section 2 details the hybrid knowledge graph construction methodology. Section 3 describes the tagging strategy and BERT-BiGRU-CRF model . Section 4 evaluates experimental results, and Section 5 concludes with future directions.

2. Hybrid knowledge graph construction methodology

KG is basically a special semantic network composed of nodes and edges, which can connect different kinds of information together to form a relational network based on the connections between things.

The construction of KG can be divided into three types: top-down, bottom-up and the combination of the two. In the top-down approach, the ontology concept layer (i.e. pattern layer) is constructed from the top to down to determine the edge of knowledge extraction, and then the entity is added to the knowledge base through the graph construction technology such as knowledge extraction. In the bottom-up approach, entities, relationships and attributes with high confidence coefficient are extracted from data sources and added to the knowledge base. Then concepts are abstracted from the down to top to complete the construction of the pattern layer. The method of combining the two is to build the pattern layer from the top to down, and then to build data layer

from the bottom to up. Through the induction and summary of newly acquired data, the entity expansion is realized based on the updated pattern layer [4].

The KG can be divided into general KG and vertical KG . General KG is not aimed at specific fields, the accuracy of knowledge is not high, and the breadth of knowledge is highlighted. The DBpedia[12], Yago[13], Freebase[14], Wikidata[15]are typical examples. Vertical KG, such as MusicBrainz[16], IMDB[17], is oriented to a specific field, emphasizing the depth of knowledge, and demanding the professionalism and accuracy of knowledge.

The KG of marine ranching equipment is vertical KG, which generally adopts the top-down construction approach. However, with the increase of the amount of data, the difficulty of updating and maintaining the graph will become more and more prominent. Therefore, this study adopts a combination of the two methods to construct the KG of marine ranching equipment, which can not only support large data quantity, but also ensure the high quality of knowledge.

The process includes data acquisition and preprocessing, knowledge modeling, knowledge extraction and data storage,which is shown as follows:
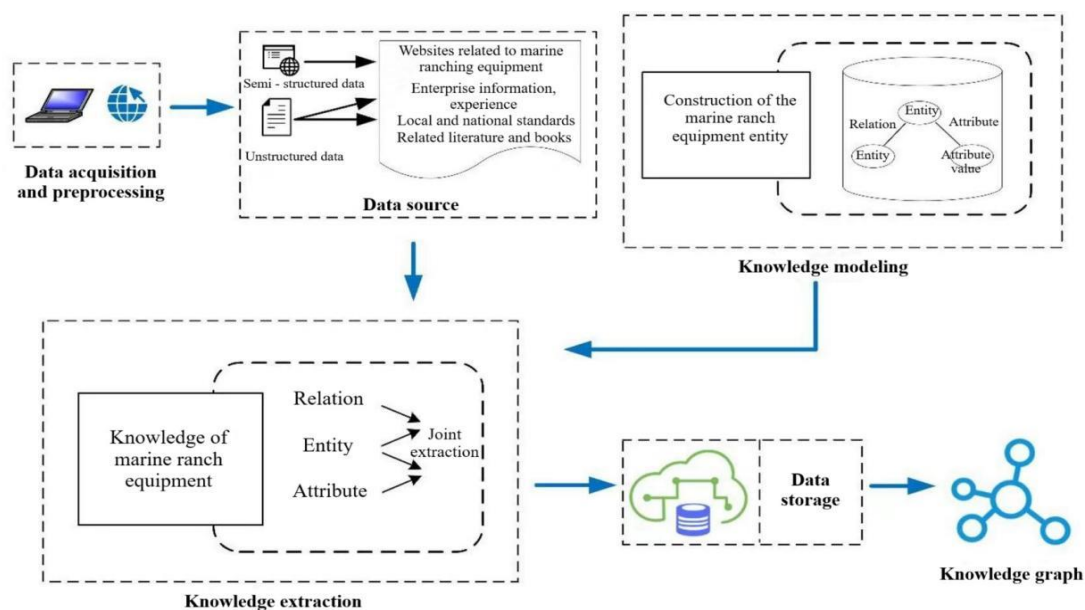
Figure 1. Construction process of marine ranching equipment KG.

2.1. Data acquisition and preprocessing

The primary data sources for constructing the KG include related websites, enterprise production records, expert interview, local and national standards, as well as relevant literature and publications. These data sources are categorized into semi-structured data and unstructured data. To ensure both the quality and quantity of acquired knowledge, distinct acquisition and preprocessing methods are implemented for different data types.

For semi-structured data from websites, the most popular data acquisition technique is web crawling. After comprehensively evaluating popular web crawling frameworks—such as Scrapy, PySpider, Crawley, and Portia.The Scrapy was selected due to its advantages in stability, speed, scalability, modular structure, and low inter-module coupling [18].However, raw HTML documents obtained often contain irrelevant content and redundant information, which may compromise the quality and efficiency of subsequent knowledge extraction. Therefore, preprocessing is essential to perform data cleaning and format standardization, ensuring the reliability of knowledge sources for graph construction. The specific workflow is illustrated in Figure 2.

Figure 2. Preprocessing workflow diagram.

Step 1: Collect and analyze the related websites, and specify the crawling target;

Step 2: According to the website structure, write corresponding scripts to obtain raw HTML format documents;

Step 3: Use regular expressions to clean HTML document, remove the advertising, labels, etc;

Step 4: Write a format conversion script and combine with certain manual review (such as clearing spaces, duplicate content, etc.) to sort out the JSON text research document in {" key ":" value ", "key" : [value]} format.

Unstructured data such as enterprise data, expert interview, relevant literature and books can be divided into electronic text data and paper text data.For the electronic text data, text parsing method is used to obtain it. The paper text data is obtained by OCR text recognition.In order to facilitate the unified processing of subsequent data, combined with manual audit, the obtained data is cleaned and the format is converted to obtain the JSON file in the same format as above.

Figure 3 shows the processed data, which does not contain any irrelevant content and is subject to certain rules. The "value" containing one value is stored as a string, and the "value" containing multiple values is stored as an array. Each piece of data represents a type of marine ranching and its associated attributes and attribute values.

Figure 3. The processed data.

2.2. Knowledge Modeling

Knowledge modeling is not only the foundation and preparation work for the construction of KG, but also the premise for the complete construction of valuable KG. It can effectively organize and utilize useful knowledge in massive information to build a unified knowledge model that is convenient for computer processing [19]. Ontology is a modeling tool that describes domain concepts, which can ensure that the graph has good structure and redundancy. Therefore, this study adopts the ontology-based modeling method to build the pattern layer of the graph.

The graph in this study is vertical KG.It has higher professional knowledge and accuracy requirements, therefore adopting the method of top-down, manual building construct ontology. At present, common manual ontology construction methods include seven-step method [20], skeleton method [21], METHONTOLOGY method [22], etc. The seven-step method is the most widely used method at present. It is an iterative ontology modeling method, which is mainly used for the construction of domain ontology. The advantages are that it has detailed step description and strong operability. The steps are as follows: Determine the domain of the ontology; Consider whether existing ontologies can be reused; List ontology key items; Determine the types and structure of types;

Determine the attributes of the types; Identify the characteristics of the attribute; Create an instance. For ontology design, there is no absolutely correct domain ontology construction method,but the most suitable method for a certain application scenario. By referring to ontology construction methods in other fields, this study combined with the application scenario of marine ranching equipment field, optimized the existing seven-step method, and finally obtained the construction process of marine ranching equipment field ontology, as shown in Figure 4.
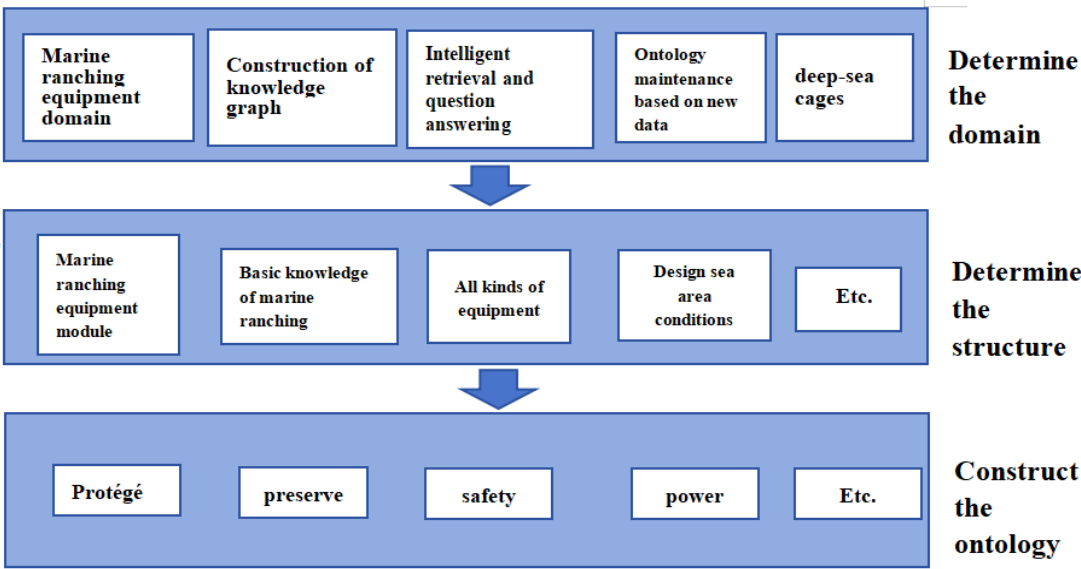
Figure 4. Construction process of marine ranching equipment ontology.

2.2.1. Determine the domain of the ontology

Protégé[23] is an open source ontology editing tool developed by Stanford University Biomedical Information Research Center based on Java programming language.When using protégé to build an ontology, the primary task is to determine the domain of the ontology, that is, we should clarify what the domain covered by the ontology is, what its purpose is, what scenarios it will be applied to and how to maintain it. In this study, the ontology covers the field of marine ranching equipment, which is mainly used for the construction of KG . The data in the ontology will be used for intelligent retrieval and question-answering. The main maintenance method is to update classes, relationships and attributes based on the induction and summary of new data. In addition, listing key concepts and terms in the domain gives users and builders a clearer understanding of the entire ontology database. Some of the key concepts and terms are shown in Table 1.

Table 1. Key concepts and terms.

Core Concepts	Domain terminology
	Jinghai No.1
Deep-sea cages	Chinesebao No.1
	Changjing No.1
	Salinity sensor
Environmental monitoring equipment	Flow velocity sensor
	Remote-Control intelligent winch
	Emergency generator
Energy equipment	Diesel generator
	UPS uninterruptible power supply
	Radar

Navigation equipment	Manual horn
	Radar transponder
	AIS (Automatic Identification System)
	Life jacket
Security equipment	Life raft
	CO$_2$ fire extinguisher

2.2.2. Determine the structure and related elements

It is necessary to consider and design the concept, attribute and relation of marine ranching equipment ontology comprehensively. According to field investigation and relevant papers, marine ranching equipment can be divided into the following five equipment modules: multi-functional platform, deep-sea cages, marine ranching observatory, sea fishing boat and large-scale aquaculture vessel[1]. Based on the five equipment modules and the existing data content and characteristics, the remaining four parent concepts are determined: Various types of equipment, marine design criteria, positioning methodology and principal dimensions, among which all kinds of equipment include seven sub-concepts such as security equipment, energy equipment, aquaculture equipment and navigation equipment. In order to further increase the amount of graph data and adapt to the current situation that marine ranching tends to be intelligent , this study further enriches the ontology database and adds two parent concepts, marine ranching equipment knowledge and marine ranching construction. The concept of marine ranching equipment knowledge mainly includes the knowledge of existing demonstration areas of marine ranching and the introduction of various types of marine ranching. The concept of marine ranching construction refers to local standards, including monitoring and evaluation, layout and distribution of construction norms. Figure 5 shows the structure.
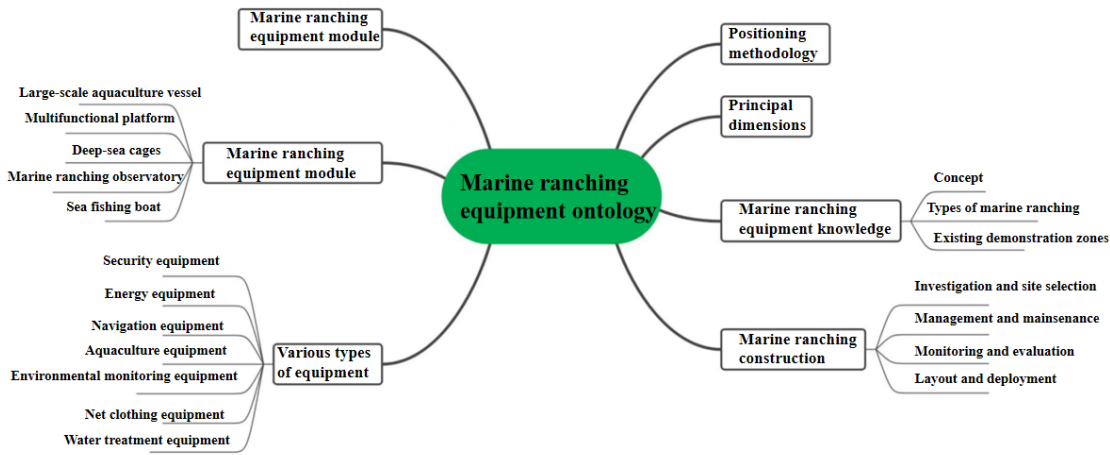
Figure 5. Seven core concepts of the marine ranching.

For each concept has different features and corresponding data, attributes need to be defined. Table 2 lists some attributes of the marine ranching equipment ontology.

Table 2. Attributes of the ontology of marine ranching equipment(sample).

Attribute/label	Domain	Range
Introduction	Multi-functional platform, large-scale aquaculture vessel and other	String

	equipment module	
Company	Deep-sea cages, sea fishing boat and other equipment module	String
Function	Multi-functional platform,deep-sea cages and other equipment module	String
Preserve	Aquaculture equipment, energy equipment and other equipment	String
Note	Deep-sea cages,water treatment equipment and other equipment	String
Put	Aquaculture equipment , security equipment and other equipment	String
Define	Marine ranching areas(e.g., equipped marine ranching type)	String
Facility	Marine ranching areas(e.g., equipped marine ranching type)	String

In the above ontology structure, in addition to the upper and lower relation between concepts, there are also certain semantic relations among entities contained in sub-concepts. For example, there is a use_condition relationship between the multi-functional platform (" Geng Hai No. 1 ") and marine design criteria (" designed water depth (10m) "), and there is aquaculture between the deep-sea cages (" Jing Hai No. 1 ") and the aquaculture equipment (" Automatic bait feeder (1 set) ") And the category of relationships is the same as the range. Table 3 shows the semantic relationships of the marine ranching equipment body based on the concepts designed in the previous section.

Table 3. Semantic relationships of marine ranching equipment.

Relationship	Relationship description	Domain	Range/ Relationship category
Safety	The relationship among multi-functional platform, deep-sea cages and security equipment.	Multi-functional platform, deep-sea cages	Security equipment
Conduction	The relationship among multi-functional platform, deep-sea cages, and navigation equipment	Multi-functional platform, deep-sea cages	Navigation equipment
Netting	The relationship between deep-sea cages and net clothing equipment	Deep-sea cages	Net clothing equipment
Power	The relationship between large-scale aquaculture vessel and energy equipment	Large-scale aquaculture vessel	Energy equipment
Aquaculture	The relationship between deep-sea cages and aquaculture equipment	Deep-sea cages	Aquaculture equipment
Positioning_mode	The relationship among multi-functional platform, deep-sea cages, and positioning methodology	Multifunctional platforms, deep-sea cages	Positioning methodology
Use_condition	The relationship among large-scale aquaculture vessel, marine ranching observatory and marine	Large-scale aquaculture vessel, marine ranching,	Marine design criteria

	design criteria	observatory	
Principal_dimension	The relationship among multi-functional platform, sea fishing boat and principle dimension	Multi-functional platform, sea fishing boat	Principle dimension

2.2.3. Marine ranching equipment ontology construction

Finally, the ontology modeling tool protégé was used to build the ontology of marine ranching equipment.The above defined concepts, related elements and some examples were added to complete the knowledge modeling of the graph, as shown in Figure 6.
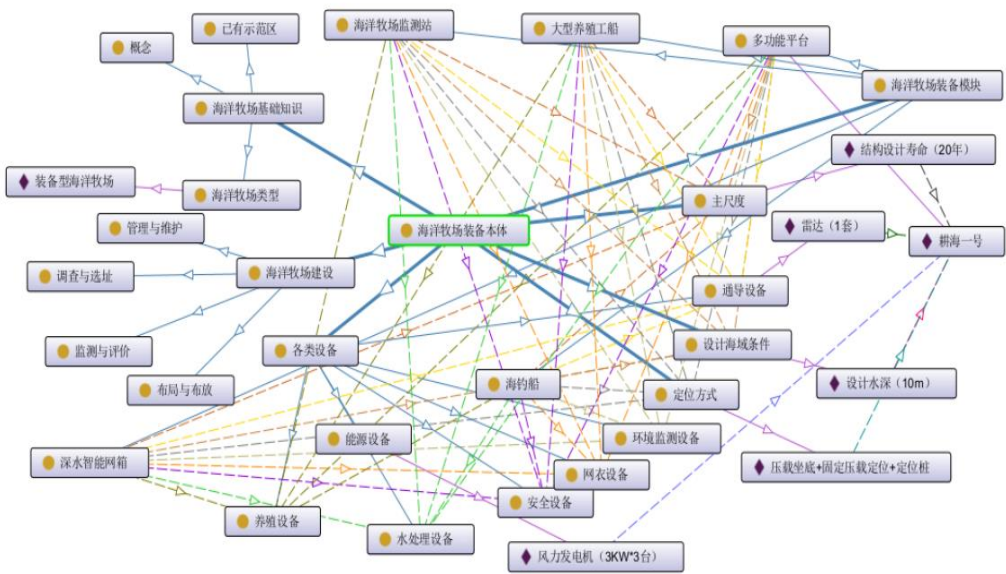

Figure 6. Marine ranching equipment knowledge ontology database**.**

2.3. Knowledge extraction

Knowledge extraction aims to extract triple data from different types of acquired data, so as to provide necessary knowledge for the construction of KG. It is divided into three categories: entity, relation and attribute extraction.Entity extraction is the most basic and key step in knowledge extraction. Deep learning-based methods are currently the most popular. In the field of knowledge extraction, common deep learning models include BiGRU-CRF[24], BiLSTM-CRF[25]. In this study, it needs to identify marine ranching equipment entities in the marine ranching equipment text, such as "Geng Hai No. 1" (marine ranching equipment module instance), "equipped marine ranching type" (marine ranching type instance), etc. Relation extraction is to extract the relationship between entities on the basis of entity extraction. Attribute extraction refers to the extraction of entity attribute information such as ""equipped marine ranching type" and "Geng Hai No. 1", and generally regards attributes as the relationship between entities and attribute values for extraction [26].

The extracted objects includes semi-structured data and unstructured data. Extraction methods for different types of data is not the same. For semi-structured data, the method of rule-based scripting is used to complete the joint extraction of entity attributes. For unstructured data, the method based on deep learning model is used to complete the joint extraction of entity relations. The specific extraction model is described in Section 3.

2.4. Knowledge Storage

Knowledge storage should consider the application scenario and data scale , then choose the appropriate storage mode to store the structured knowledge in the database, which can realize the efficient management and analysis of data. At present, knowledge storage can be divided into two kinds according to the storage structure: table-based knowledge storage and graph-based knowledge storage. After comprehensive analysis, this study adopts the graph-based knowledge storage .

Presently, the predominant graph database systems encompass HyperGraphDB, OrientDB, and Neo4j. Neo4j is the most popular among them , which can store and query entities, attributes and relationships in the knowledge graph, and can also support applications to operate and analyze the knowledge graph.[27]which is classified as a property graph within the realm of graph databases.And it is fundamentally structured around four core components: labels, nodes, relations, and attributes . The effects and description object represented by each constituent are delineated in Table 4. In contrast to alternative graph database systems, Neo4j boasts superior scalability, the capacity to accommodate millions of nodes on standard hardware configurations, the availability of the Cypher query language, and compatibility with a multitude of popular programming languages. Consequently, Neo4j has been chosen as the database platform for the storage and maintenance of the graph structure in the present study.

Table 4. Neo4j basic elements description**.**

Neo4j basic element	Effect	Description object
Labels	Describe ontology concept	Multi-functional platform,security equipment and other ontology concepts
Nodes	Describe entity	Examples of radar, diesel generator, etc
Relations	Describe the relationships between entities	Security equipment and other relations
Attributes	Describe attributes of entities and relationships	Attributes **s**uch as name, model, and operation mode

This study is based on the py2neo library in Python, and through script writing, it facilitates the batch import of triples, such as (entity, relationship, entity) and (entity, attribute, attribute value). The Neo4j-based knowledge graph encapsulated 2,153 nodes and 3,872 edges.,Figure 7 shows partial content.

In Figure 7, where in nodes distinguished by varying colors correspond to disparate conceptual instances, while the varying edges interlinking these nodes signify the interrelational aspects. Owing to the openness of the KG and the commendable scalability of the Neo4j database, it is anticipated the KG established in this research can be systematically enriched and augmented via Cypher query language statements. This will, in turn, provide a robust foundation for subsequent applications in equipment fault diagnosis, maintenance planning, and policy formulation.
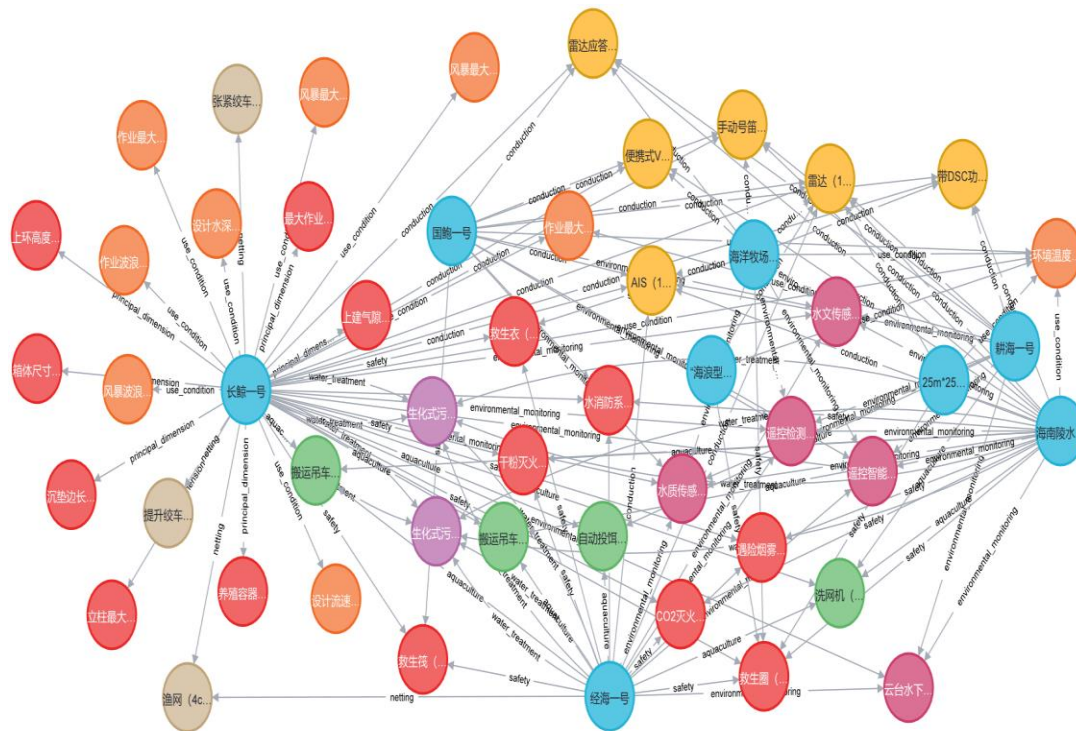
Figure 7. Knowledge graph of marine ranching equipment based on Neo4j**.**

3. Joint Extraction of Knowledge in the Field of marine ranching equipment

Knowledge extraction aims to extract triplet data from different types of data obtained, so as to provide necessary knowledge for the construction of KG.As mentioned above, for semi-structured data, rule-based joint extraction is used to complete the joint extraction, for unstructured data, the method based on deep learning model is used to complete the joint extraction .

3.1. Rule-based joint extraction of entity attributes

As depicted in Figure 3, the dataset analyzed in the preceding section is characterized as semi-structured data, adhering to specific rules. For certain data entries, the initial key-value pair enclosed within each set of curly braces denotes the category to which the entity pertains, as well as the entity's name. Subsequent key-value pairs are archived in the format "attribute": "attribute value," with each pair pertaining to the entity identified in the initial key-value pair. Empirical validation has confirmed that this structured rule facilitates the extraction of entity-attribute-attribute value triples (e.g., marine ranching equipment entity-attribute 1-attribute value 1; marine ranching equipment entity-attribute 2-attribute value 2; ...; marine ranching equipment entity-attribute n-attribute value n).

In order to enhance the presentation and utility of the KG, the current study introduces a method to normalize multi-valued attributes into entities. A segment of the KG is illustrated in Figure 8.
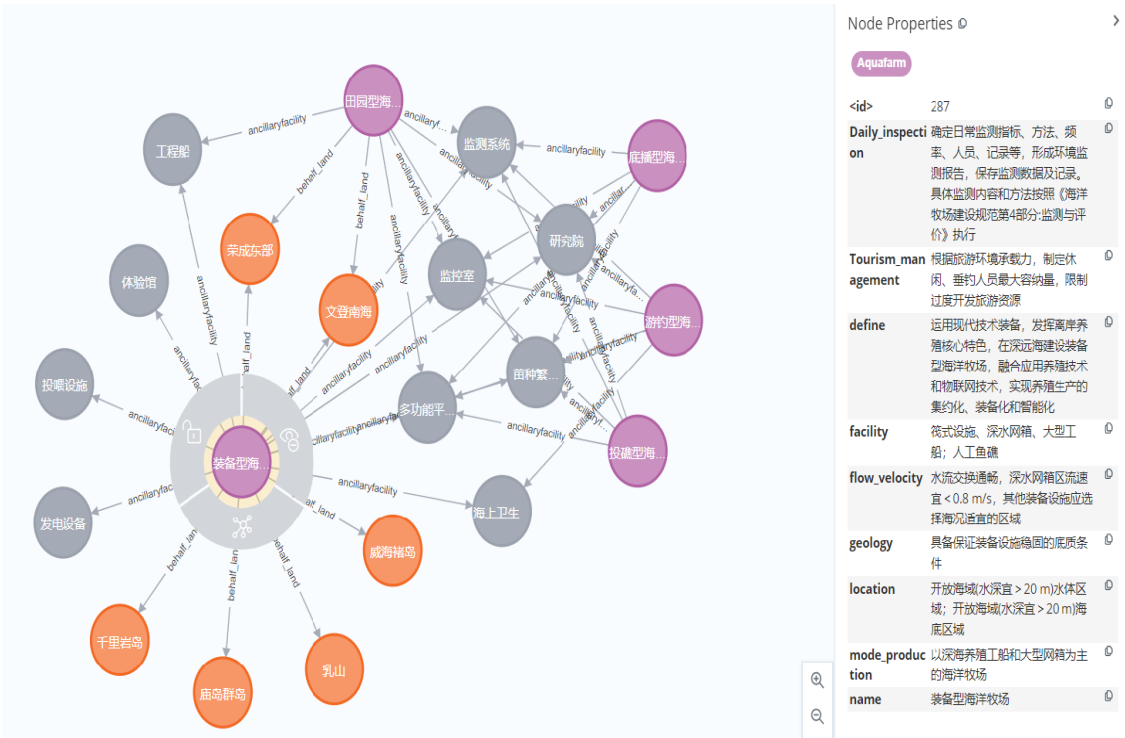
Figure 8. Example of knowledge storage in semi-structured text.

3.2. Joint Entity Relation Extraction Based on Deep Learning Models

3.2.1. The innovative TE+SE+R$_i$+BMESO tagging strategy

Upon examining the "JSON" dataset that was previously processed, it has been ascertained that the "value" field comprises a substantial portion of unstructured textual data, which concurrently harbors numerous cryptic interconnections among entities. For example, within the "value" segment of the "introduction" (profile) attribute for "Jinghai No.1", there are intricate entity relationships pertaining to principal dimensions, aquaculture equipment, and marine design criteria.

Drawing upon the comprehensive analysis of the marine ranching equipment corpus, in conjunction with the interrelations delineated within the model layer, several distinctive attributes have been elucidated: (1) The extraction tasks for this iteration are unanimously centered around the conceptual entity of the marine ranching equipment module, thereby designating the marine ranching equipment module entity as the theme entity within the extracted triples; (2) The relations between the marine ranching equipment module entity and other entities, as well as the categorization of these other entities, remain consistent. The identification of the entity types facilitates the determination of their respective relationships; (3) In a sentence, it is feasible to encounter multiple relationships between the marine ranching equipment module entity and other diverse entities.

Drawing upon the preceding analytical insights, the current study introduces an innovative tagging strategy, designated as TE+SE+R$_i$+BMESO, which is specifically tailored for the marine ranching equipment corpus. The study employs the BERT-BiGRU-CRF entity extraction model to concurrently identify and extract inter-entity relationships. Within the context of this tagging schema, the marine ranching equipment module entity is denoted as the theme entity, represented by TE. Entities that interact with the marine ranching equipment module entity are denoted by SE_ R$_i$, with SE denoting the secondary entity and R$_i$ indicating the category of the i-th secondary entity SE$_i$, which corresponds to the relationship type linking the theme entity to SE$_i$. The BMESO sequence labeling approach is utilized, with the detailed connotations of each label delineated in Table 5.

Table 5. TE+SE+Ri+BMESO tagging strategy and description of tags.

Tag	Description
TE	Theme entity
R_i	Relationship type between theme entity and secondary entity (SE_i)
B-TE	Beginning character of theme entity
M-TE	Middle character of theme entity
E-TE	Ending character of theme entity
S-TE	Theme entity of single character
B-SE_R_i	Beginning character of secondary entity (SE_i)
M-SE_R_i	Middle character of secondary entity (SE_i)
E-SE_R_i	Ending character of secondary entity (SE_i)
S-SE_R_i	Secondary entity (SE_i) of single character
O	Other non-named entity characters

3.2.2. The specific structure and working principle of the BERT-BiGRU-CRF model

The BERT model[18], a groundbreaking advancement in natural language processing (NLP) in recent years, demonstrates exceptional performance in text representation and semantic understanding. The BERT-BiGRU-CRF architecture comprises three layers: a BERT layer, a Bidirectional Gated Recurrent Unit (BiGRU) layer, and a Conditional Random Field (CRF) layer. Leveraging the Transformer architecture, BERT enables parallel computation, significantly enhancing training and inference efficiency compared to conventional models based on Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN), particularly when processing large-scale datasets. The BiGRU layer excels in sequence modeling by autonomously capturing contextual information, offering advantages over Bidirectional Long Short-Term Memory (BiLSTM) networks, including a simpler architecture, fewer parameters, and faster computational speed. The CRF layer incorporates global sequence dependencies by calculating optimal label sequences, thereby effectively utilizing inter-label relationships. Currently, the BERT-BiGRU-CRF framework has been widely adopted for Named Entity Recognition (NER) across diverse domains. For instance, Wang Yuquan[28] et al. enhanced the precision, recall, and F1-scores in geotechnical engineering text NER by proposing a pre-trained BERT-BiGRU-CRF language model. Similarly,Ma Wenxiang [29] et al addressed challenges in Chinese resume NER, such as character-level polysemy and representation limitations, through their integrated BERT-BiGRU-CRF approach.The overall model is shown in the figure 9.
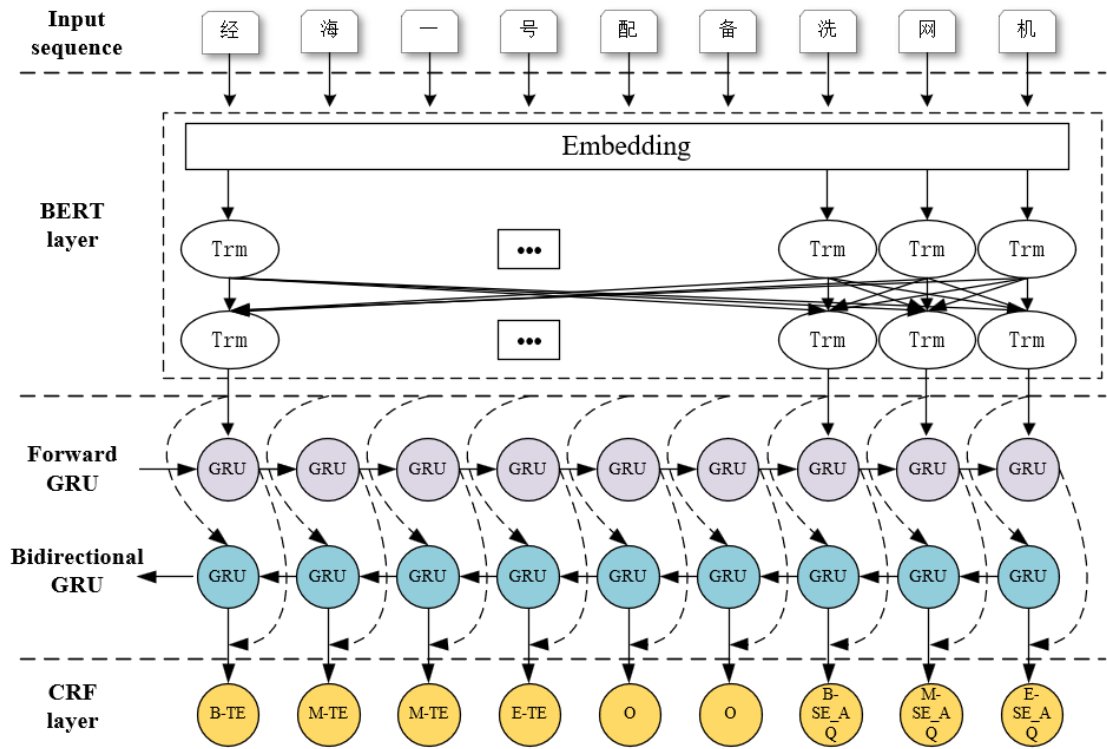
Figure 9. BERT-BiGRU-CRF model structure**.**

(1)BERT layer

BERT is a context-based word embedding model, boasts enhanced capabilities in deriving nuanced semantic attributes from textual data compared to its non-contextual predecessors like word2vec. Its proficiency in addressing the challenge of polysemy through contextual analysis is rooted in its core architecture the bidirectional transformer encoding structure. The intricacies of which are delineated in Figure 10.
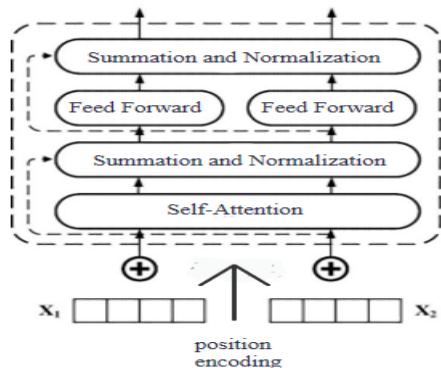

Figure 10. Transformer encoding structure**.**

The execution of the BERT layer predominantly encompasses two pivotal components: the representation of input data and pre-training procedures. The representation of input data pertains to the transformation of the data into a format compatible with BERT's input requirements. Each character within the input data is   the aggregate of token embeddings, segment embeddings, and position embeddings, as shown in Figure 11[30].
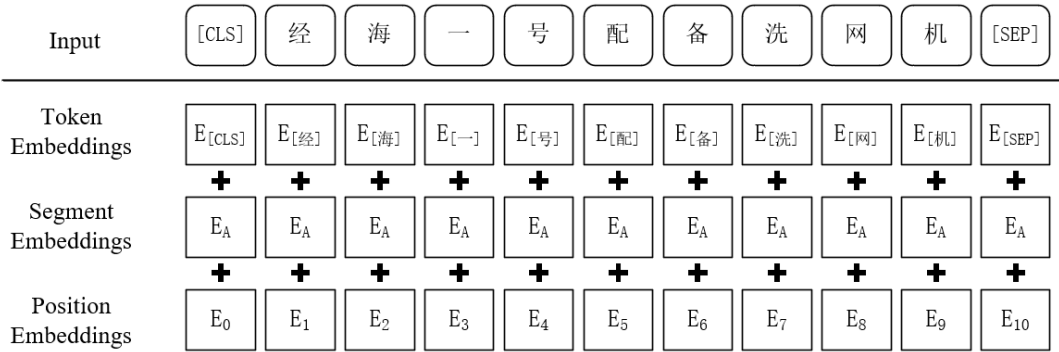
Figure 11. Input representation of BERT model.

BERT is pre-trained based on two major tasks: "Masked Language Model" (MLM) and "Next Sentence Prediction" (NSP). Through the simultaneous training of these two tasks, it can better extract word-level and sentence-level features of the text, obtaining token embeddings that contain more semantic information.

(2)BIGRU layer

The GRU is an alternative iteration of the Recurrent Neural Network (RNN) architecture. In contrast to the tripartite gate mechanism of the LSTM, the GRU has been streamlined to incorporate just two gates: the reset gate and the update gate, as depicted in Figure 12. To concurrently acquire contextual insights, the current study has developed a Bidirectional Recurrent Neural Network (BiGRU) network, where in the fundamental building block consists of both forward and reversed GRU elements.
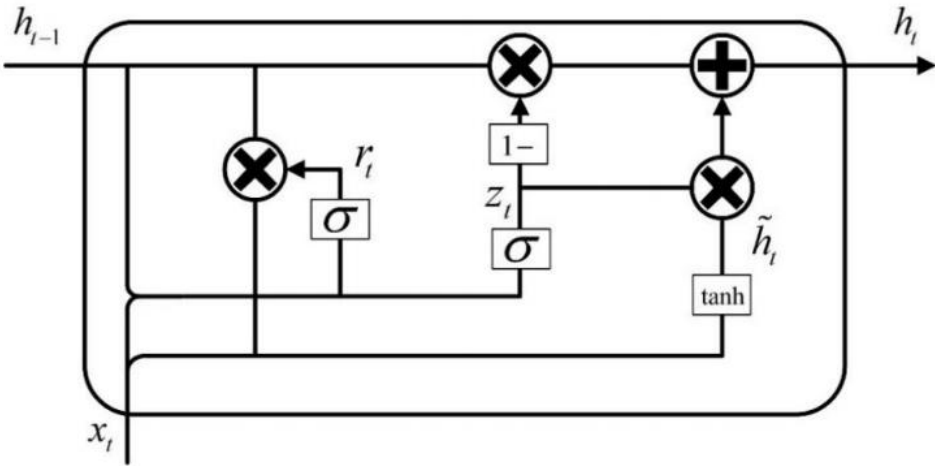

Figure12. GRU gate model.

In Figure12, the variable x_t denotes the input data at the current instance, whereas h_t signifies the output at the current instance, and h_{t-1} denotes the output at the previous instance. Concurrently, the reset gate r_t and the update gate z_t operate synergistically to regulate the previous hidden state h_{t-1} and to facilitate its transition into the new hidden state h_t. The reset gate combines h_{t-1} and x_t, producing a matrix r_t, the elements of which range from 0 to 1, and which dictate the extent of information from the previous output h_{t-1} that is to be discarded. A value closer to zero indicates a higher degree of information erasure from h_{t-1}. The specific formula is as follows.

$$r_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_r) \qquad\qquad （1）$$

The update gate combines h_{t-1} and x_t to control how much information from the previous step's output h_{t-1} is retained. The smaller the value, the more information from h_{t-1} is retained, and the less information from the current step is retained. The specific formula is as follows.

$$z_t = \sigma(Wz \cdot [h_{t-1}, x_t] + b_z) \tag{2}$$

The candidate memory $\overline{h_t}$ consists of two components: one part is the current input data x_t, and the other part is the output h_{t-1} from the previous moment, determined by the reset gate r_t. When r_t equals 0, it signifies a complete erasure of past information. The specific formula is as follows.

$$\overline{h_t} = \tanh(Wz \cdot [h_{t-1}, x_t] + b_h) \tag{3}$$

The matrices W_r, W_b and W_z denote the weight matrices, while b_r, b_z, and b_h represent the biases. The ultimate output is determined by the update gate, and the formula for this computation is as follows .

$$h_t = (1 - z_t) * h_{t-1} + z_t * \overline{h_t} \tag{4}$$

(3)CRF layer

In the domain of named entity recognition, it is imperative to acknowledge the existence of interdependencies among the labels. For example, the label "B-TE" is invariably followed by the proscription of the label "M-SE_AQ". Nonetheless, the BIGRU model inadvertently conforms to a strategy of selecting the label with the highest probability as the predicted outcome, without due consideration for the inter-label constraints. In response to this challenge, the current study introduces the integration of a CRF (Conditional Random Field) layer. During the label prediction phase, this CRF layer meticulously evaluates both the individual probabilities of each label and the transition probabilities derived from the training corpus, effectively mitigating the likelihood of illicit labels and enhancing the precision of the predictive outcomes.

Suppose the input sequence is X = {x1, x2, x3, ..., xn}, and the output label sequence is y = {y1, y2, y3, ..., yn}. The score calculation formula is as follows:

$$score(x, y) = \sum_{i=0}^{n} A_{yi, yi+1} + \sum_{i=1}^{n} pi, yi \tag{5}$$

In the formula, A is a transfer matrix of size (k + 2) $\times$ (k + 2), where $A_{i,j}$ represents the score of label i transferring to label j. P is the output matrix of the BIGRU layer, with a size of n$\times$k, where n indicates the sentence length and k represents the number of labels. $p_{i,j}$ denotes the score of the i-th word being marked as the j-th label.In order to derive the probabilities corresponding to all potential tag sequence scores, the softmax function is employed. The mathematical formula is delineated as follows:

$$P(y \mid X) = \frac{e^{Score(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{Score(X,y)}} \tag{6}$$

In the given formula, Y_X denotes the comprehensive set of all feasible label sequences corresponding to the input sequence X, with y~ representing the actual label sequence. Subsequently, a logarithmic transformation is applied to both sides of the equation, followed by the application of the Viterbi algorithm for decoding, thus identifying the sequence with the highest scoring value. The detailed computational formula is delineated as follows:

$$\log p(y \mid x) = score(x, y) - \log(\sum_{\tilde{y} \in Y_X} e^{Score(X,y)}) \tag{7}$$

$$y^* = ar g \max_{\tilde{y} \in Y_X} Score(x, \tilde{y}) \tag{8}$$

4. Results and Analysis

4.1. System testing environment

In the context of this investigative endeavor, the experimental procedures were executed utilizing the Python and PyTorch frameworks. The corresponding software and hardware configuration are delineated in Table 6.

Table 6. Hardware and software configuration.

Configuration environment	Parameter
Operation system	Windows 10 64-bit
CPU	AMD Ryzen 7 5800X
GPU	NVIDIA GeForce RTX 3090（24G）
Memory	32G
Python	3.9.13
PyTorch	1.13.1

This investigation assesses the efficacy of the model by employing three pivotal performance metrics indigenous to the domain of knowledge extraction, viz., precision (P), recall (R), and the F1 score. The corresponding mathematical formula are delineated as follows:

$$P = \frac{TP}{TP + FP} * 100\%$$

$$(9)$$

$$R = \frac{TP}{TP + FN} * 100\%$$

$$（10）$$

$$F_1 = \frac{2 \times P \times R}{P + R} * 100\%$$

$$（11）$$

In the given formula, "TP" — the number of entities that have been accurately identified; "FP" — the number of entities that have been inaccurately identified; "FN" — the number of entities that have eluded detection.

4.2. Experimental data and parameters

In this study, the dataset consists of 1,456 annotated sentences specifically related to marine ranching equipment. To garner more profound insights, the dataset was subdivided utilizing the cross-validation methodology, adhering to a proportion of 8:1:1, which yielded a training corpus of 1,164 sentences, a validation corpus of 146 sentences, and a test corpus of 146 sentences. Following an array of parameter fine-tuning trainings, the optimal parameter settings are outlined in Table 7.

Table 7. Experimental parameter settings.

Parameter	Parameter value
batch_size	64
epoch	30
seq_max_len	128
learning_rate	0.01
GRU hidden state dimension	300

4.3. Experimental results

In order to substantiate the superiority of the model developed within the scope of this research, a comparative analysis was conducted against the prevalent algorithmic model, BERT-BiLSTM-CRF, in the domain of knowledge extraction. The detailed outcomes of these experiments are delineated in Table.8, while the recognition efficacy pertaining to diverse entities is graphically showed in Figure 13.

Table 8. Comparison of model effects.

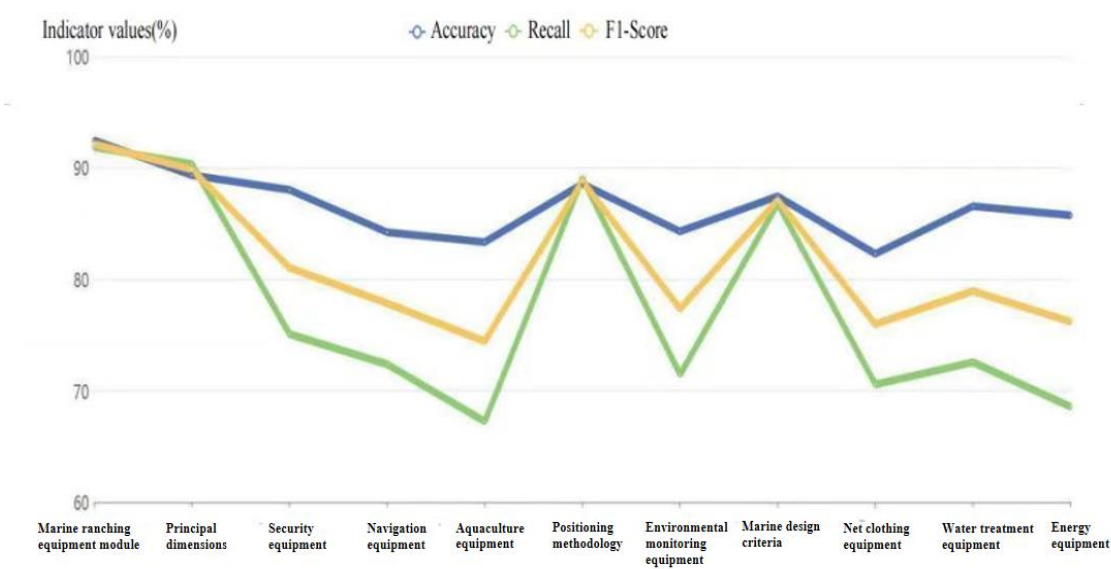| Model	Precision P/%	Recall R/%	F1 score /%
BERT-BiLSTM-CRF	85.37	75.32	80.03
BERT-BiGRU-CRF	86.58	77.82	81.97

Figure 13. Entity recognition result**.**

As illustrated in Table 8, experimental results demonstrated superior performance over BERT-BiLSTM-CRF, achieving 86.58% precision, 77.82% recall, and 81.97% F1-score(1.94% improvement, $p < 0.05$) . In contrast to the BiLSTM model,,the BiGRU model boasts a reduced parameter count, which not only enhances model performance but also accelerates the training process. The utilization of BiGRU as the encoding layer is deemed more appropriate for the text-based entity recognition task specific to marine ranching equipment.

Furthermore, an analysis of Figure 13 and associated data reveals that the F1 score for the marine ranching equipment module entity is the highest at 92.17%. Notably, entities such as marine design criteria, positioning methodology, and principal dimensions exhibit considerable F1 scores at 87.12%, 88.81%, and 89.85% respectively. The presence of diverse nomenclature for various equipment entities—such as "batch feeder", "bait dispenser", and "GQ48902", all referring to the "feeding machine" in the context of aquaculture equipment—leads to an increased prevalence of unrecognizable entities and manifests as a relatively low recall rate. Consequently, this impacts the overall performance of the model, yielding a lower recall rate and F1 score.

5. Conclusions and Prospects

In essence,this study not only pioneers the first structured KG framework for marine ranching equipment, but also offers a transferable methodology for vertical domain knowledge extraction, yielded successful outcomes. The present study employs a hierarchical, top-down methodology to establish the model layer of marine ranching equipment , culminating in the formulation of an ontological framework for the marine ranching equipment KG. Subsequently, a bottom-up method is enacted to develop the corresponding data layer, facilitating the comprehensive acquisition of marine ranching equipment data and the subsequent extraction. Thereafter, the BERT-BiGRU-CRF model is used to accomplish the joint extraction of entity relationships within the marine ranching

equipment domain. Ultimately, the graph data is stored within the Neo4j database. The conclusions are summarized as follows:

The Neo4j-based knowledge graph encapsulated 2,153 nodes and 3,872 edges, enabling scalable visualization and dynamic updates. Experimental results demonstrated superior performance over BERT-BiLSTM-CRF, achieving 86.58% precision, 77.82% recall, and 81.97% F1-score(1.94% improvement, $p < 0.05$)

Currently, there is still room for improvement in the KG, and it will be further deepened and expanded from the following aspects:

In the initial phase, it is imperative to address the dynamic and continuous evolution of marine ranching equipment, which is characterized by rapid advancement and an ever-expanding body of knowledge. The capacity to capture data in real-time and facilitate the dynamic updating of the marine ranching equipment KG represents a pivotal challenge for future endeavors.

Furthermore, the KG established in this research is presented solely in textual format. It is proposed that future initiatives explore the integration of visual or auditory data, such as images, audio, and video, to construct a multifaceted KG that enhances connectivity with the tangible world.

Finally, the KG of marine ranching equipment can be combined with a large language model. ChatGPT is a hot research topic in the field of intelligent question-answering, which can quickly understand and answer users' questions in various general fields, but for vertical fields such as marine ranching equipment, its reliability and professionalism are still insufficient. Combining the KG of marine ranching equipment with the large language model can give full play to the advantages of both, providing users with personalized equipment selection suggestions, and realizing more intelligent and efficient knowledge management and application.

References

1. Zhao, Z.; Ding, J.; Ji, Y.; et al. Exploration and Practice of Modern Information Technology and Engineering Equipment in Marine Ranching Construction .*China Fisheries*.2020,(04), 33-37.
2. Shen, W.; HU, Q.; Shen, T.; et al.Research and experiment of intelligent floating fish accumulating equipment for ocean ranching. *Journal of Shanghai Ocean University*.2016,25(02): 314-320.
3. Shi, J. Intelligent equipment technology for offshore cage culture.*Beijing: Ocean Press.*2018, 1-159.
4. Huang, H.; Yu, J.; Liao, X.; et al.Knowledge graph research review.*Computer Systems & Applications*.2019,28(06),1-12.doi:10.15888/j.cnki.csa.0069.
5. Dieter, F.; et al. Introduction: what is a knowledge graph?.2020.doi:10.1007/978-3-030-37439-6_1
6. Ren, T.Research and Application of OPC UA Information Model Based on Knowledge Graph.Postgraduate,Zhejiang University , China, 2021.doi:10.27461/d.cnki.gzjdx.2021.001377.
7. Gu, X.; Bao, J.; Lü, C.; et al.Assembly Semantic Information Modeling Based on Knowledge Graph.*Aeronautical manufacturing technology*.2021,64(04),74-81.doi:10.16080/j.issn1671-833x.2021.04.074.

8. Huet, A.; Segonds, F.; Pinquie, R.; et al. Context-aware cognitivedesign assistant:Implementation and study of design rules recommendations. *Advanced Engineering Informatics*.2021,50, 101419.doi:10.1016/j.aei.2021.101419

9. Hao, Q.; Hu, S. Smart design on the flexible gear based on knowledge graph. *Journal of 29 Physics: Conference Series (Vol. 1885; No. 5; p. 052021).* doi:10.1088/1742-6596/1885/5/052021

10. Zhou, B.; Hua, B.; Gu, X.; Lu, Y.; Peng, T.; Zheng, Y. An end-to-end tabular information-oriented causality event evolutionary knowledge graph for manufacturing documents.*Advanced Engineering Informatics*.2021,50,101441.doi:10.1016/j.aei.2021.101441

11. Hao, X. Research on prediction and treatment method of workshop production anomaly based on LSTM.Postgraduate, Harbin University of Science and Technology, China, 2020.doi:10.27063/d.cnki.ghlgu.2020.000540.

12. Bizer C.; Lehmann J.; Kobilarov G.; et al. DBpedia - A crystallization point for the Web of Data.*Web Semantics Science Services & Agents on the World Wide Web*. 2009, 7(3):154-165.doi:10.1016/j.websem.2009.07.002.

13. Suchanek F M.; Kasneci G.; Weikum G. Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web, ACM, 2007, 697-706.

14. Bollacker K D.; Cook R P.; Tufts P. Freebase: A shared database of structured general human knowledge. In: Proceedings of the 22nd National Conference on Artificial Intelligence, 2007, 2: 1962–1963.

15. Vrandečić D. Wikidata: a new platform for collaborative data collection. Proceedin gs of the 21st International Conference on World Wide Web. ACM, 2012: 1063-1064.

16. IMDB Official. IMDB[EB/OL]. [2016-02-27]. http://www. imdb.com.

17. MetaBrainz Foundation. Musicbrainz[EB/OL]. [2016-06- 06].http://musicbrainz.org/.

18. Sun, Y.Design and Implementation of Web Crawler System Based on Scrapy Framework. Postgraduate.Beijing Jiaotong University.China.2019.

19. Yun, W. ;Zhang, X.;Li, Z.; et al. Knowledge modeling: A survey of processes and techniques.*International Journal of Intelligent Systems*, 2021(3).doi: 10.1002/int.2235

20. Noy, N. Ontology development 101: A guide to creating your first ontology.2001.

21. Uschold, M.; Gruninger, M. Ontologies: Principles; methods and applications. *The knowledge engineering review*.1996.11(2); 93-136.doi:10.1017/S0269888900007797.

22. Jones, D.; Bench-Capon, T.;Visser, P. Methodologies for ontology development. *Capon*. 1998:62--75.doi:http://dx.doi.org/.

23. Knublauch, H.; Fergerson, R .; Noy, N.; et al.The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications//*International Semantic Web Conference.Springer, Berlin, Heidelberg*,2004.doi:10.1007/978-3-540-30475-3_17.

24. Xu C.; Wang F.; Han J.; et al.Exploiting multiple embeddings for Chinese named entity recognition//*Proceedings of the 28th ACM International Conference on Information a nd Knowledge Management*,2019:2269-2272 .

25. Huang Z.; Xu W.; Yu K .Bidirectional LSTM-CRF Models for Sequence Tagging.*Computer Science*.2015.doi:10.48550/arXiv.1508.01991.

26. Ma, Z.;Ni, R.;Yu, K. The latest progress, key technologies and challenges of knowledge graph. *Journal of Engineering Sciences*.2020,42(10):1254-1266.doi:10.13374/j.issn2095-9389.2020.02.28.001.

27. Dou, J.; Qin, J.; Jin, Z.; et al. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage.*Journal of Visual Languages & Computing*. 2018, 48(OCT.):19-28.doi:10.1016/j.jvlc.2018.06.005.

28. Wang, Q.; Li, Z.; Tu, Z.; et al. Geotechnical Named Entity Recognition Based on BERT-BiGRU-CRF Model. *Earth Science*. 2023,48(08): 3137-3150.doi:org/10.3799/dqkx.2022.462

29. Ma, W.; Liao, T.; Zhang, S.; et al.Named Entity Recognition of Chinese Electronic Resume Based on BERT-BiGRU-CRF. *Journal of Yancheng Institute of Technology(Natural Science Edition)*.2022,35(03): 41-47.doi:10.16018/j.cnki.cn32-1650/n.202203008.

30. Wang B.; Wang A.; Chen F.; et al.Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*.2019, *8*. doi:10.1017/ATSIP.2019.12