IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 February 2023, accepted 3 March 2023, date of publication 8 March 2023, date of current version 21 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3254132

APPLIED RESEARCH

The Construction of Knowledge Graphs in the Aviation Assembly Domain Based on a Joint Knowledge Extraction Model

PEIFENG LIU[1], LU QIAN[2], XINGWEI ZHAO[1], AND BO TAO[1], (Member, IEEE)

[1]State Key Laboratory of Digital Manufacturing Equipment and Technology, Department of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2]School of Transportation and Logistics Engineering, Wuhan University of Technology (WHUT), Wuhan 430063, China

Corresponding author: Xingwei Zhao (zhaoxingwei@hust.edu.cn)

This work was supported by the National Science Foundation of China under Grant 52275020, Grant 62293510, and Grant 91948301.

ABSTRACT The aviation assembly domain, which is a complex system, involves the multi-dimensional information of parts, processes, tools, plants and operation projects. In order to assist the knowledge management from natural language text in the aircraft manufacturing process, this paper proposes the corresponding ontology scheme and the joint knowledge extraction model, which is necessary for construct the knowledge graph in the aviation assembly domain. The model is able to automated end-to-end construct knowledge graph. The proposed model, which is based on reinforcement learning approach and a novel labeling scheme, takes the constraint relationships between entities and relations as an important identification basis. The model does not rely on manual feature setting, while it greatly reduces the training cost. The proposed joint knowledge extraction model was testified from the practical scenarios of the general assembly and component assembly. The experimental results showed that the proposed model showed an excellent performance in the aviation assembly domain, with the F1-score of 89.71% for entities, the F1-score of 91.27% for relations, and the overall average F1-score of 82.41%. Based on the superior performance of the model, the knowledge graph of the general assembly and component assembly, which included 1, 308 pairs of triples composed of five kinds of entities and three kinds of relations, was further constructed in the aviation assembly domain.

INDEX TERMS Intelligent manufacturing, aviation assembly, knowledge graph, knowledge extraction, joint learning.

I. INTRODUCTION

Knowledge Graphs (KG) have recently garnered significant attention from both industry and academia in scenarios [1]. By organizing human knowledge into structured information, KG is able to use for many intelligent applications as crucial resources [2]. In 2012, KG, which were constructed of precisely interlink data, were firstly introduced by Google as the next generation of intelligent semantic search engine technology [3]. KG are consist of nodes and edges, where the nodes represent either concepts, concrete objects, data, or information resources about them, and the edges represent semantic relationships between the nodes [4]. Through representing human knowledge as a computer-processable data structure, KG thus are widely used in many professional domain [5]. However, compared with the general domain, the construction of KG in the specialized professional domain is still a challenging work. Several major problems which should be paid attention to are brought forward, such as the insufficient specialized corpus, the unclear structure of KG, the robustness of knowledge extraction model.

Hence, in order to solve the above problems, this paper presents a full-process scheme for automatically constructing KG from natural language texts in a new professional domain.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiwu Li.

In this paper, the aviation assembly domain is considered as an example of specific professional one, and the KG of general assembly and component assembly is constructed.

It has an important theoretical significance and practical value to study the KG in the aviation manufacturing field. The related aviation assembly technology is one of the core technologies in the aviation manufacturing field. In the process of the aviation assembly, high requirements for the professional knowledge and professional ability of the assembly engineers are necessary, however training qualified and skilled assembly engineering personnels often require a lot of time, manpower and material resources [6]. Thus, constructing a corresponding KG, which is a more comprehensive and effective knowledge support tool and decision aid tool in the assembly process, is proposed to improve the assembly efficiency and reduce the assembly error rate.

Inspired by the approach of joint information extraction [7] and reinforcement learning (RL) [8], this paper proposes a model of joint knowledge extraction in the aviation assembly domain. The proposed framework takes into account the transfer probabilities between entities and relations in the domain knowledge graph. Meanwhile, by employing a novel labeling scheme, the framework models the entity and relation extraction as a joint sequence labeling task. The model makes full use of the correlation information between entities and relationships. Thus, the model, which is an end-to-end knowledge extraction method without manual pre-constructed features, is used to construct KG automatically.

The contribution of this paper is mainly manifested in the following three aspects:

1) Combining the professional knowledge of the experts in the field of aviation assembly, rules for entity classification are established by multiplexing the ontology consensus of the existing relevant domain knowledge base [9]. At the same time, the relation classification rules between entities are established. Thus the structure of triples in the aviation assembly domain is established as well. Based on the previous work, a novel labeling scheme is introduced to jointly extract the entities and relations as a sequence labeling problem. Therefore the annotated corpus of the aviation assembly domain is constructed. The corpus data source of this paper is from the Aviation Manufacturing Engineering Manual [10].
2) In this paper, we propose a knowledge graph construction of the aviation assembly domain, which is based on a joint knowledge extraction model. Moreover, the experiment based on the practical scenarios of the general assembly and component assembly has proven that the model shows excellent performance in the aviation assembly domain. The weighted average F1-score is 82.41% with the F1-score of 89.71% for entities and the F1-score of 91.27% for relations. The recognition performances always stay high level without relying on manual pre-constructed features. The proposed model can extract the core value data from the massive data of the actual scenario in the aviation assembly domain, so that it also provides a technical basis for the subsequent construction of the knowledge graph in the aviation assembly domain.
3) Based on the superior performance of the model, the KG of general assembly and component assembly, which includes 1, 308 pairs of triples composed of 5 kinds of entities and 3 kinds of relations, is further constructed in the aviation assembly domain.

To sum up, in this paper, new data is present from a the professional book and rather reliable, a structure of KG in the aviation assembly domain is proposed, a joint knowledge extraction model is described, and finally the KG of general assembly and component assembly is constructed.

II. RELATED WORK

To construct a KG is to extract data as computer-processable triplets $(e1, R, e2)$ from the unstructured texts, where $e1$ is the agent entity, $e2$ is the object entity, and R represents the semantic relation between $e1$ and $e2$ [11]. Thus entities and relations extraction are important tasks to construct KG [12]. To realize the automation of knowledge graph construction, it is necessary to establish the technical model of automatic entities and relations extraction. With the development of sequential statistical algorithms and deep neural network algorithm models, the technical path of automated knowledge graph construction has become clear. In this case, the task of building a knowledge graph can be further decomposed into the named entity recognition task (NER task) [5] and the relation classification task (RC task) [13]. The purpose of NER task is to extract entities, and the purpose of RC task is to extract their relations. NER task and RC task can obtain good identification results by using the sequential statistical algorithms and the deep neural network algorithm models, which play an important role in the construction of knowledge graph.

The two tasks are often performed separately as in pipeline engineering. Firstly the entity pairs are extracted, and then the corresponding relationships of entity pairs are extracted. Therefore, the relevant knowledge extraction method in such schemes is called the pipelined scheme [14]. In such scheme, each sub-tasks are flexible to deal with, however, the relevance of two sub-tasks are often unobserved [15], thus it loses a lot of relevant contextual semantic information. Furthermore, since the RC task is often performed after the NER task, such errors generated in the NER task are propagated into the RC task, resulting in a geometric-level error [16]. In many existing technical frameworks, NER and RC are performed as two independent tasks with two mutually independent models. It not only consumes a lot of training resources, but also brings the problem of error accumulation [17].

To address these problems, recently, more and more joint learning models are proposed to jointly extract entities and relations. Joint learning model can learn more contextual information and avoid cascading errors [16]. Joint learning via the joint model of NER and RC extracts entities and relations simultaneously, thus the corresponding

entity-relation-entity triplets are achieved directly. Early in the study of joint learning models, most existing joint methods are feature based structured systems [18], [19], [20], which need complicated feature engineering [15]. With the development of joint learning, the joint end-to-end model without manual setting features becomes the more efficient scheme. Due to the differences in modelling objects, the main technical paths of the joint learning models can be divided into parameter sharing schemes [21] and sequence annotation schemes [17].

The parameter sharing schemes are to model both entities and relations. Such methods share parameters in the encoding layer, then the final decision is made by two networks separately in order to detect entities and relation types. During training, by using the backward propagation algorithm, the shared parameters of the two sub-task encoding layers are updated simultaneously, so that the two sub-tasks are interdependent. Finally, the global optimal parameters are found through a lot of training, thus the best performing entity and relation joint extraction model is achieved.

The sequence annotation schemes are to directly model the entity-relation-entity triples. To realize the purpose of joint learning in the sequence annotation schemes, the two NER and RC subtasks are completely transformed into a sequence annotation problem through new annotation strategies. The proposed method solves the problem of information redundancy in the parameter sharing scheme well, and directly obtains the entity-relation-entity triples through an end-to-end model without manually setting a large number of features.

Miwa and Bansal [22] proposed a joint learning scheme. Compared with the pipeline scheme, the joint learning scheme is able to simultaneously extract the entity pairs and the corresponding relationships, thus the problems existing in the pipeline scheme can be very well solved at the same time. However, the scheme still requires the manual pre-constructed features. Zheng et al. [23] proposed a new bidirectional encoder-decoder model, which encodes the input sentence through BiLSTM, and corresponding two decoders were designed for the entities recognition and relations classification. Giannis et al. [24] proposed a joint neural model, which uses a Conditional Random Fields (CRF) layer for the named entity recognition task and regards the relation extraction task as a multi-head selection problem. The scheme is able to carry out the task of knowledge extraction jointly and end-to-end, without the process of manual pre-constructed features.

To improve the robustness of model, the RL was introduced [25], [26]. Takanobu et al. [27] designed a hierarchical RL model to extract relations and entities. Chen and Teng [11] proposed a model, which regarded the removal of noisy data as a RL process. The purpose of RL agents was to determine whether the candidate corpus should be removed from the training dataset. The proposed scheme demonstrated excellent performance in a constantly changing training environment.

III. METHODOLOGY

The process framework for constructing the knowledge graph is specifically shown in the Figure 1. Firstly, conduct the Optical Character Recognition (OCR) of the corresponding chapter of the book. Then clean the data including eliminating the pictures and tables, consequently the book is turn into pure text. Later, the pure text, which is regraded as the corpus, is imported into the secondary developed human-computer interaction tool to annotate entities and relations. The annotated corpus is classified into the training, the validation, and the test sets. The joint knowledge extraction model designed in this paper is trained on the training set and is used to predict the test set corpus afterwards. The triples are extracted, thus the proposed algorithm is able to assess the validity of the model. Structured knowledge data is stored in the graph database after knowledge fusion, and finally forms the knowledge graph in the aviation assembly domain.

A. THE FRAMEWORK OF KNOWLEDGE GRAPH IN THE AVIATION ASSEMBLY DOMAIN

1) THE DATA SOURCE OF CORPUS IN THE AVIATION ASSEMBLY DOMAIN

The datasets applied in this paper are from the book ''Aviation Manufacturing Engineering Manual: Aircraft Assembly''. This book is used for craft personnel engaged in aircraft assembly and designers engaged in aircraft engineering design. It can be used as a reference material for technical workers in aircraft component assembly, general assembly, aircraft commissioning, test flight and cable manufacturing, as well as a teaching reference book for teachers and students in aviation colleges [10]. Therefore, it has a very high degree of professionalism and credibility. This paper annotates the contents of Article 3, General Aircraft General Assembly and Commissioning, including Chapter 9, General Assembly Process Design as the corpus data source for general assembly, and Chapter 10, Part Docking as the corpus data source for component assembly.

2) THE TEXT CORPUS ANALYSIS OF KNOWLEDGE GRAPH IN THE AVIATION ASSEMBLY DOMAIN

The knowledge graph of the general domain is very mature. Wikipedia has opened up the TB-level corpus, providing a lot of data base for the construction of the knowledge graph. Compared with general domain knowledge graphs, data on professional domain knowledge graphs requires not only conceptual knowledge, but also a data corpus that can reflect deeper relationships, therefor it needs to be collected and annotated separately. In some professional domains, there are already large and standardized public corpus, such as the MSRA [28] which is annotated and published by Microsoft Research Asia, the media social corpus [29], and the talent resume corpus [30]. The corresponding knowledge graph is established on those corpus.

In order to construct the aviation assembly knowledge graph, the Aviation Assembly (AA) corpus is first

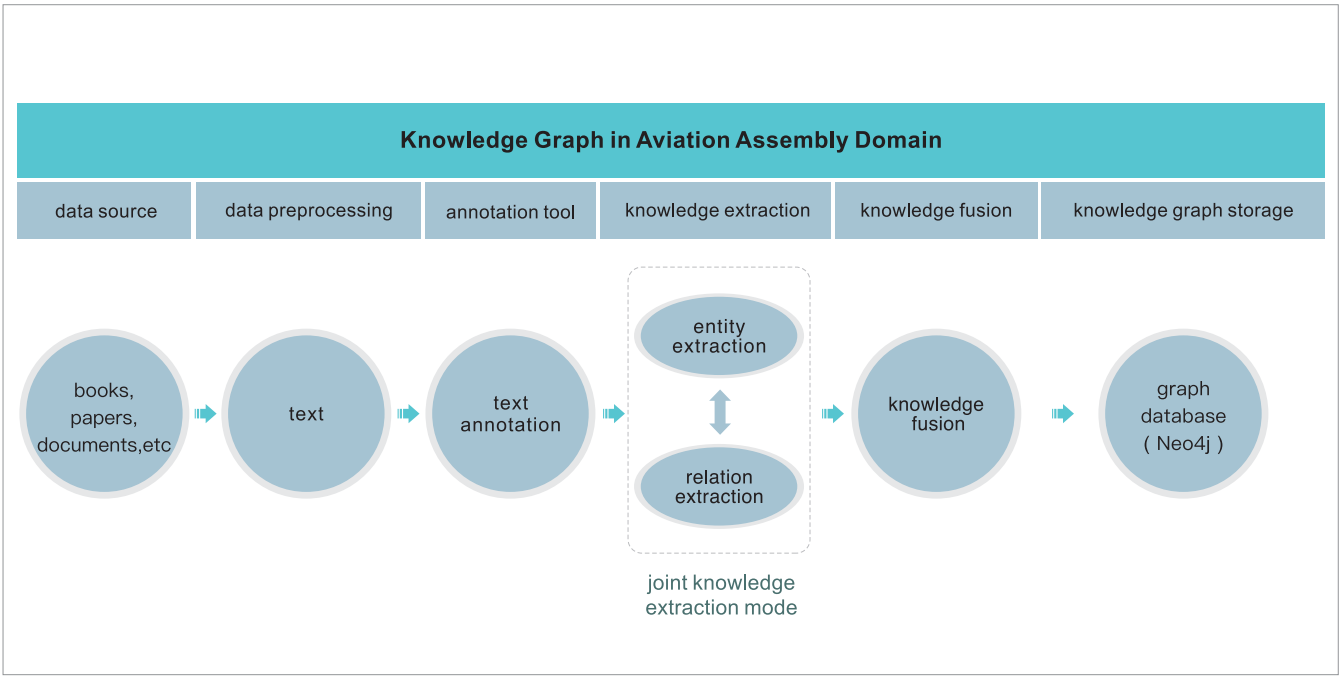

FIGURE 1. Construction process of knowledge graphs in the aviation assembly domain.

established. In the test datasets, the entity characters account for 32.15% of all characters, and this ratio is also called ER rate. While, the ER ratio to MSRA is 16.11%.

The characteristics of the knowledge graph in the aviation assembly domain are as follows:

1) The quantity of corpus is small.
2) High professional degree of corpus.
3) Due to the first two characteristics, the noise content is low, and the entity ratio (ER ratio) is high.
4) The named entity boundary in the corpus is blurred.

Combined with the above characteristics and compared with the general domain, the aviation assembly domain knowledge graph puts forward higher requirements for the joint knowledge extraction model, which can be summarized as follows:

1) The classification of entities and relations in the aviation assembly domain is more complex and specialized. The naming format of entities and relations in general domain is relatively clear, standardized and unified. For example, entity categories can be divided into people, positions, places, organizations, etc., and relation categories can be divided into friends, relatives, colleagues, etc. While, entities and relation classification in the field of aviation assembly needs to combine expert knowledge and engineering scenarios. It is not only necessary to ensure that the extracted entities and relations can fully restore the assembly process, but also keep the possibility of large-scale expansion. In this paper, firstly, we effectively establish the entity classification of the aviation assembly domain by multiplexing the ontology consensus of the existing relevant domain knowledge base. Then the relation classification is defined according to the relationship between the entity classification.
2) Data sources in the aviation assembly domain are highly unstructured. In the construction of the corpus in the professional fields, the data sources are mainly focused on the relevant professional books, papers, and internal enterprise documents and national standards. Such sources of data provide information often in unstructured texts.
3) The fuzzy entity boundary exists in the aviation assembly domain. For example, ''radar calibration'' can be collectively defined as the operation entity, or ''Radar'' as the component entity and ''Calibration'' as the operation entity, respectively; For another example: the ''engine short cabin'' can be defined as a single component entity overall, or it can be defined as the ''engine'' and the ''short cabin'' as two component entities respectively. In the process of boundary definition, the principle of maintaining the semantic integrity of the entity is adopted. Such as ''Radar calibration'' and ''Engine short cabin'' are defined as a complete entity. At the same time, different entities classification needs to have certain linguistic characteristics, especially in the essential features of the words. The words' essential features are helpful to the annotation and training process. For example, component entities, facility entities, and tool entities are nouns, and operation entities and operation step entities are mostly combination phrases of nouns + verbs.

4) The construction process of the corpus in the aviation assembly domain requires the high performance of the algorithm. The corpus in the professional domain often needs to be built from scratch. In order to make the construction process more efficient, the designed algorithm needs to consider the operation speed and the recognition effect comprehensively. The construction of corpus is a process of human-computer interaction, and machine annotation and manual audit are often conducted alternately. Therefore, the machine recognition effect of the algorithm has certain requirements on the speed of algorithm training to ensure the tempo of human reading and fluency of the whole human-computer interaction process.

Based on the characteristics and difficulties of the corpus in the aviation assembly domain, the performance of the model puts forward higher requirements.

3) THE ARCHITECTURE OF ENTITIES AND RELATIONS IN THE AVIATION ASSEMBLY DOMAIN

Combined with expert knowledge in the field of aviation assembly and related books, literature and other materials, the article firstly defines the approach of entity classification. Five kinds of entities are defined, including the component entity, the facility entity, the operation entity, the step entity, and the tool entity. The defined approach of entity classification is based on multiplexing the ontology consensus [9] of existing relevant domain knowledge bases and considering the gaps in Chinese and English as well. It means that the KG constructed in this way can be relatively easy to understand and integrate with other KG. The correspondences between the corresponding ontologies of the five entities (aviation assembly ontologies) and the ontologies of existing relevant domain knowledge bases are shown in the Table 1.

TABLE 1. Comparison between the aviation assembly ontology and the existing domain ontology.

Aviation Assembly Ontology	Intermediate Engineering Ontology [9]
component	product
facility	plant
operation	process
step	process plan
tool	tool

Among the Table 1, the Intermediate Engineering Ontology (IEO) [9], which is the ontology consensus of the existing relevant domain knowledge base, bridges the gap between the top ontology (top-level ontologies) and the existing domain knowledge base.

The five kinds of entities in the aviation assembly domain are defined as follows: the components of the aircraft are defined as component entities, such as ''engine'', ''landing gear'', etc.; the corresponding facilities in the aircraft assembly are defined as facility entities, such as ''apron'', ''main plant'', etc.; the manually operated projects are defined as operation entities, such as ''radar calibration'', ''system sealing experiment'', etc.; the actual process of aircraft assembly is defined as step entities, such as ''general assembly'', ''component assembly'', etc.; and the specific tools used in the assembly process are defined as tool entities, such as ''jack'', ''power support facilities'', etc.. This entity classification method can specifically reproduce the assembly implementation details in five dimensions, namely, time (the step entity), location (the facility entity), operation project (the operation entity), tool (the tool entity), and operation object (the component entity).

After a clear entity classification mechanism, the relations between the corresponding entities can be well defined. This paper defines three types of relations, with profiles and examples (bold for solid pairs) as follows:

1) Instrument-Agency (IA): The relationship of operations and objects. Examples: the **engine (component) test run (operation)**, the **installation (operation)** of **engine (component)**, etc.
2) Component-Whole (CW): the relationship between whole and parts. Example: **power plant (component)** includes **engine (component)**, **aircraft assembly (step)** includes **total assembly (step)** and **component assembly (step)**, etc.
3) Content-In (CI): the relationship of content in time or space. Example: **engine (component) commissioning (operation)** is in the **test flight station (facility)**, **engine (component) test run (operation)** is in the **system function test (step)**, etc.

In this paper, five kinds of entities and three kinds of relations are defined in the aviation assembly domain to construct the triples of the domain knowledge graph, see the Figure 2. Based on this basis, the knowledge graph can fully restore the details and logical relationships of the assembly knowledge, that is, at which time, where, what tools to use, to which parts, and what to do.

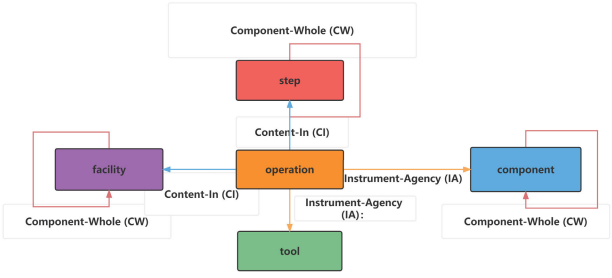

FIGURE 2. Architectural diagram of entities and relationships in the aviation assembly domain.

In the Figure 2, it is clear that transfer probabilities exist between entities and relations. For example, the relation CI only happens between operation entities and facility or step ones, and do not happen between operation entities and tool ones. Thus sequential statistical algorithms or layers should considered for the task. Furthermore, the specific transfer probabilities of the corpus in this paper are able to calculate according to tagging texts.

4) THE NOVEL TAGGING SCHEME OF KNOWLEDGE GRAPH IN THE AVIATION ASSEMBLY DOMAIN

With the novel tagging scheme, the two NER and RC subtasks are completely changed into a sequence annotation problem, in order to achieve the purpose of joint learning. The technical route has shown excellent performance in many tasks, such as the construction of knowledge graph [31] in the biomedical domain. The core of the scheme is to incorporate more semantic information into the labels.

Through the analysis of entities and relations in the aviation assembly domain, relations only exist between entities, that is, non-entities do not exist relations. Thus, entities can be divided into two categories, namely, entities with relations and entities without relations. After this analysis, the annotation scheme, which contains three parts of the semantic information, can be defined as following:

1) The location information of the characters in the entity. This paper uses the BMEO tag scheme to make character-based tags, namely, B represents the starting character of an entity, M is the middle character of an entity, E is the end character of an entity, and O is the character of a non-named entity.
2) The category information of the entity. Five entity categories are defined in the Table 1.
3) The relationship information of the entity. This paper defines 3 kinds of relations, see the Figure 2.

In this paper, the corpus data is annotated through the proposed novel tagging scheme, see the Figure 3 for an example. In order to more efficiently annotate the data and display the model effects, this paper has further developed a human-computer interaction tool, which is more suitable for progressive tasks. The developed human-computer interaction tool is based on the doccano [32] text annotation tool, see the Figure 3. The core purpose of the secondary development is to use the trained model to advanced annotate the entities and relations, therefore the annotation task is transformed into a review task.

In this paper, the corpus with 10, 640 characters are marked for the general assembly and component assembly, moreover the space characters and the punctuation marks are included. The quantity statistics of the entities and relations in the corpus are shown in the Table 2:

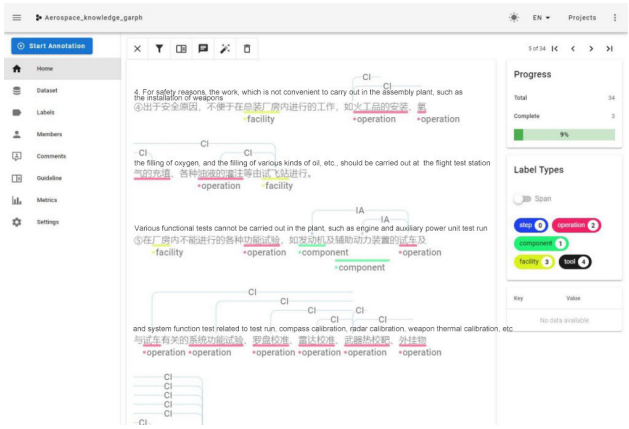

FIGURE 3. Human-computer interaction interface for natural language annotation.

TABLE 2. The quantity statistics of corpus entities and relations.

Entities	Unit/individual	Relations	Unit/individual
component	328	Instrument-Agency	254
facility	92	Component-Whole	67
operation	262	Content-In	22
step	122	—	—
tool	155	—	—

B. THE FRAMEWORK OF THE JOINT KNOWLEDGE EXTRACTION MODEL

1) OVERVIEW OF THE JOINT KNOWLEDGE EXTRACTION MODEL

Combined with the current situation of knowledge graph research in the field of aviation assembly, this paper needs to start the construction from the zero corpus. It is necessary to jointly extract the corresponding entities and relations from the self-built corpus, and finally form triples to build the corresponding knowledge graph. Therefore, this paper analyzes the performance of different kinds of models on aviation assembly corpora. Based on the performance of the models, a knowledge extraction model based on joint learning is designed. The proposed model can select and combine different statistics-based models and deep learning models in order to find a global optimal model, which balances the model robustness and high F1-score. Thus the model is able to obtain better knowledge extraction results. The proposed model can always maintain a high level in the establishment process of the aviation assembly corpus, and can effectively process the huge amount of data containing complex relationships in the aviation assembly domain.

The specific structure of the knowledge extraction model framework is shown in the Figure 4. It is divided into three layers, including the Embedding layer, the Modelling layer, and the Reinforcement layer.

This paper focuses on studying the effects of joint knowledge extraction models, so the basic method as the effect standard line in both the embedding and model layers are used. Thus, more effective observable learning effects can be achieved. In the embedding layer, the character vector training method is adopted; the model layer involves probability-based models and deep learning-based models, which includes HMM, CRF, BiLSTM, and BiLSTM + CRF models; in the reinforcement layer, the multiple models, which are established by the model layer on the validation set, are validated and used to combine the final joint knowledge extraction model based on the validation results. The role of the reinforcement layer is to implement a robust joint model with the best F1-score. To improve the robustness, it is necessary to combine more kinds of models to cope with the noise existing in the training set. Meanwhile, this combination cannot be substantially at the expense of

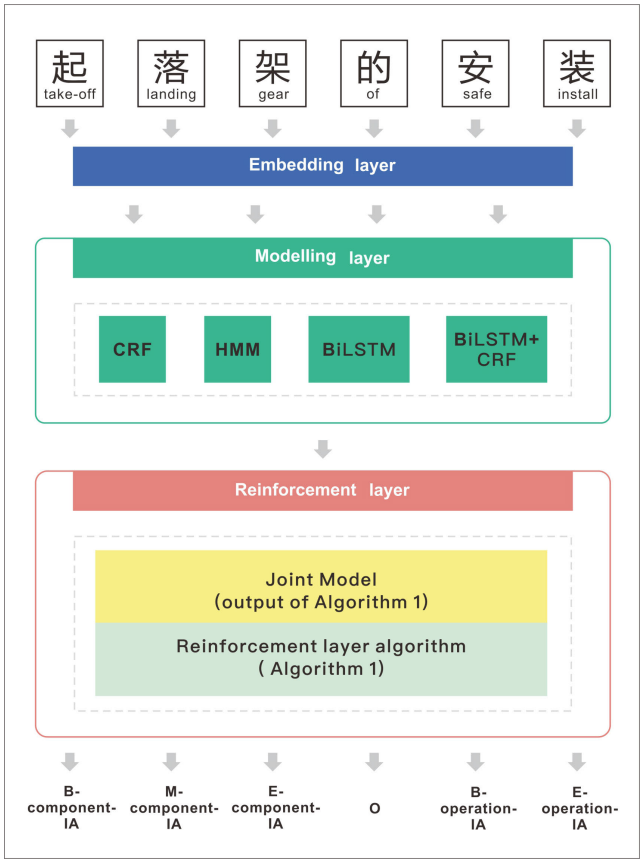

FIGURE 4. The joint knowledge extraction model framework.

F1-score. Thus, the greed index λ and the tolerance index B are introduced into the reinforcement layer according to the approach of RL. The joint model tries to combine more models (the largest greedy index λ), but only if the lost F1-score after the joint cannot exceed the tolerance index B. In this approach, the extrapolation performance of the final joint knowledge extraction model are improving. The architecture and principle of the model at each layer are shown below.

2) EMBEDDING LAYER

The task of embedding layer is to vectoring the text. In this paper, the character vector is used as the input signal, namely the character vector $c = [c_1, c_2, \cdots, c_n]$, where c_i represents the i th character vector corresponding to the i th character and the n indicates the number of character inputs. Through the compilation of the embedding layer, the new character vectors are obtained as $x^c = [x_1^c, x_2^c, \cdots, x_n^c]$, where the i th character vector x_i^c can be calculated by Equation (1):

$$x_i^c = e^c(c_j), \tag{1}$$

where e^c represents the establishing character embedding lexicon matrix based on the corpus.

3) MODELLING LAYER

The model layer is a collection of sequential annotation models. The hidden Markov model (HMM) [33] is one of them. Given an observation sequence, the character vector x^c in this paper, the final marker sequence $y = [y_1, y_2, \cdots, y_n]$ is computed to achieve the maximum conditional probability $P(Y \mid X)$. Thus the NER task is transformed to find the optimal Y^*, resulting in the maximum $P(Y \mid X)$, namely Equation (2):

$$\begin{aligned} Y^* &= \arg\max_Y P(Y \mid X) \\ &= \arg\max_Y \prod_{i=1}^{n} P(x_i \mid x_{1,i-1}, y_{1,i}) P(y_i \mid x_{1,i-1}, y_{1,i-1}) \end{aligned} \tag{2}$$

In the practice, some simplified approximation of the Equation (2) is adopted, and the optimal solution is obtained with the Veterbi algorithm.

The Conditional Random Field model (CRF) is widely favored as a simple and well-constructed model with good performance. It continuously improves in wide applications and is one of the most successful methods in named entity recognition task. The CRF [34] is an undirected graph model, whose simplest form is the linear-chain CRFs, which is well-suited for the annotation of linear data sequences. The difference with HMM model is conditional probability, which consists of two parts, one is the state probability of y_i with the input state x_i and the other is the transition probability of y_i with the previous state marked y_{i-1}. The conditional probability distribution is defined by the CRFs as the Equation (3):

$$\begin{aligned} P(y|x) = &\frac{1}{Z(x)} \\ &\times \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right), \end{aligned}$$

where

$$\begin{aligned} Z(x) = &\sum_y \\ &\times \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)\right) \end{aligned} \tag{3}$$

In the Equation (3), t_k and s_l are characteristic function, λ_k and μ_l are the corresponding weights. $Z(x)$ is the normalization factor, which is the total probabilities of all possible output sequence. The problem is similar with HMM, the NER task is translated to find the optimal Y^*, resulting in the maximum $P(Y \mid X)$. The Veterbi algorithm is also applied to obtain the optimal solution after making a simplified approximation of the Equation (3).

Long Short-Term Memory (LSTM) neural networks is a special RNN that makes up for the deficiency of the

traditional RNN. It can forget useless information while capturing long-distance important sequence information, and is therefore very suitable for NER tasks. Usually a LSTM cell contains forgetting gates, input gates, and output gates, which control the proportion of the forgotten information and passed information to the next time step.

With the input character vector $x^c = [x_1^c, x_2^c, \cdots, x_n^c]$, Bidirectional LSTM (BiLSTM) gets a hidden left to right vector $\overrightarrow{h_j^c} = \left[\overrightarrow{h_1^c}, \overrightarrow{h_2^c}, \cdots, \overrightarrow{h_n^c}\right]$ and a hidden vector from right to left $\overleftarrow{h_j^c} = \left[\overleftarrow{h_1^c}, \overleftarrow{h_2^c}, \cdots, \overleftarrow{h_n^c}\right]$. Thus the hidden vector for each character is represented as following Equation (4):

$$h_i^c = \left[\overrightarrow{h_j^c}; \overleftarrow{h_j^c}\right] \tag{4}$$

Finally, the tag sequences y corresponding to the h_i^c are able to obtain by decoding.

The BiLSTM + CRF [35], [36] model adds the CRF layer to the BiLSTM model, namely the h_i^c is used as the input to the CRF layer, and finally the tag sequence y is obtained. The BiLSTM + CRF model, when compared to the BiLSTM model, the correlation between adjacent labels can be taken into account, and the joint model yields a more accurate label sequence. Such as in the BMEO annotation system, M must appear after B, E must appear if M appears, etc.. The CRF layer can translate these problems into conditional probabilities, which contributes better results. Of course, this method takes up more computing resources. The Figure 5 shows the structural diagram of the BiLSTM + CRF model with the character inputs in this paper.

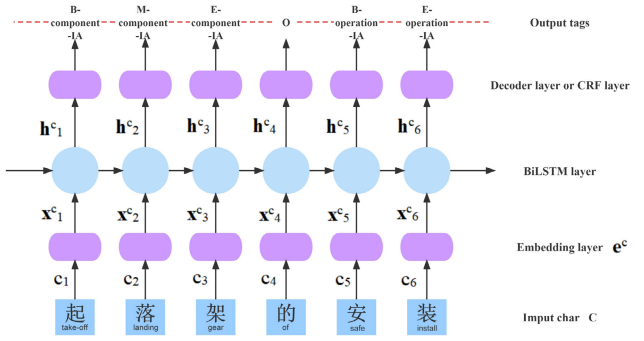

FIGURE 5. The basic structure of the Character-based BiLSTM+CRF model.

4) REINFORCEMENT LAYER

The purpose of the reinforcement layer is to screen and combine different kinds of models according to the RL approach, so that find the global optimal scheme between the robustness and higher F1-score of the knowledge extraction model, meanwhile, better knowledge extraction result can be obtained. The selected joint knowledge extraction model is used to strengthen the final prediction results as shown in the Algorithm 1, where the greed index λ, is the number of selected and combined models. The initial value of λ is the number of all models in the model layer (this paper is set to 4), indicating that the framework wants to combine the most models in order to improve the robustness of the joint knowledge extraction model. The tolerance index B is the maximum tolerate decline of F1-score by using a new model, and therefore $0 < B < 1$.

Algorithm 1 The Reinforcement Layer Algorithm of Continual Learning Framework

Input: F1 score of all models in the modelling layer, $f1 = [f1\_score\_1, \ldots, f1\_score\_m]^T$, where m is the number of models.

Output: Joint the chosen models and predicted tags

```
Require: set parameters λ and B
Sort f1 as descending order and F1 = f1[0]
while λ > 1 do
    Joint first λ models and calculate its joint F1 as F1_temp
    if F1 − F1_temp <= B then
        Break
    else
        λ = λ − 1
    end if
end while
return the joint λ corresponding models and using the joint model to predict tags
end
```

Here the joint F1-score of the joint model is the final joint prediction results for each character. Specifically, the joint prediction results of each character are separately predicted by each model in the joint model, and the most of the prediction results are taken as the final joint prediction result. If there is no majority of the results, the prediction result with the best F1-score is taken as the final joint prediction result.

The Algorithm 1 tries to combine the most models for the task, only if the F1-score lost by the joint model cannot exceed the tolerance index B. A joint model, which combines λ models, is finally obtained and strengthens the prediction results. Thus, the proposed framework can automatically screen and combine models that perform stronger in the current corpus without requiring human intervention.

IV. EXPERIMENTS

A. THE ESTABLISHMENT OF A JOINT KNOWLEDGE EXTRACTION MODEL

1) EXPERIMENTAL SETTING

The development environment of the framework is Windows10, the system type is 64-bit OS, the CPU is Intel Core $i5 - 7500@3.40$ GHz, the memory is 16GB, and no GPU is used. The developed software is python 3.8.8. The graph database uses Neo4j of version 4.3.4.

In the model layer, the model of the present paper is based on the torch with version 1.11.0. The number of hidden state

layers of the LSTM is 128, the dimension of the word-vector is 128, the number of training epoch is 30, the batch-size is 64 and the learning rate is 0.001. The model is trained by using the stochastic gradient descent algorithm. The CRF model is performed by using sklearn_crfsuite with version 0.3.6. In the reinforcement layer, the initial greed index λ is 4, and the tolerance index B is 0.01.

2) CORPUS INFORMATION

The Table 3 shows the corpus information statistics, including spaces and punctuation marks. The Aviation Assembly corpus, AA, is the corpus of the total assembly and component assembly. Among them, the validation set and the test set of AA are independent, and its selection needs to fully ensure the randomness of the corpus, while also take into account the distribution characteristics of the entity types.

TABLE 3. Corpus information statistics.

Dataset	Training/char	Validation/char	Test/char	Scheme
AA	10,318	903	576	BMEO

When dividing the corpus into training, validation and test set, the application scenario should be consider. Since in this paper, the annotation and hierarchical learning are conducted in units of a book page, the number of characters in the test set is between 500 and 1000 as the range of characters on one page of the book. Meanwhile, the corresponding validation set should be matched. The data of the validation and the test set should be randomly, however, five different entity types and three different relation types should appear simultaneously in order to ensure the generalization ability of the model and the effectiveness of the test results.

In the test set, the quantity statistics of the entities and relations are shown in the Table 4:

TABLE 4. The quantity statistics of the entities and relations.

Categories	The number of entities or relations/individual
component	28
facility	2
operation	15
step	7
tool	9
Instrument-Agency	23
Component-Whole	8
Content-In	8

3) THE EXPERIMENTAL RESULTS AND THE COMPARISON OF THE JOINT KNOWLEDGE EXTRACTION MODEL IN THE AVIATION ASSEMBLY CORPUS

The comparative experiments carry out on the aviation assembly corpus, and the experimental results of the joint knowledge extraction model and other classical models are shown in Table 5:

The calculation rules of this result are as follows: the positive samples are the ones with all the correct annotation information, that is, the three parts of the annotation scheme described in Section III-A4 of this article, namely, the location information of the characters in the entity, the category information of the entity and the relationship information of the entity, are all correct. The joint knowledge extraction model on the AA corpus finally is composed of the main CRF model and the supplemented BiLSTM + CRF model.

TABLE 5. The comparative results on the AA corpus.

Model types	Precision/%	Recall/%	F1-score/%
HMM	79.56	76.51	78.01
CRF	83.21	83.81	83.51
BiLSTM	79.81	81.49	80.64
BiLSTM+CRF	82.81	82.92	82.86
Joint	83.67	84.34	84.00

According to the Table 5, the Figure 6 can be obtained.

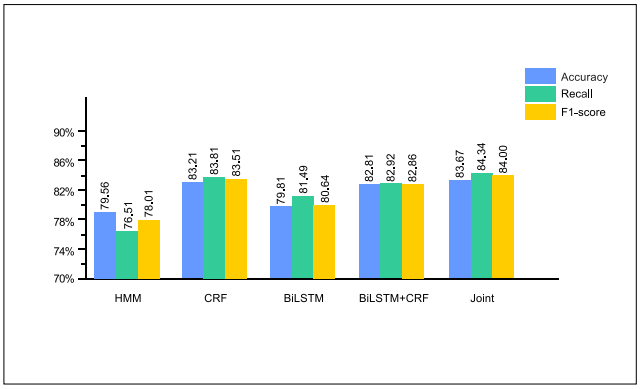

FIGURE 6. The performance comparison of the models.

It can be intuitively seen from the Figure 6 that the performance of the joint knowledge extraction model exceeds any single model, thus the advantages of the joint knowledge extraction model are proven. This result also proves that the statistics-based models achieve better results within the constrained framework of entities and relations. Professional domain knowledge graphs are often multiplexing ontologies with the empirically established structure of entities and relations, therefore, the model proposed in this paper can be well used in the construction process of professional knowledge graph.

The prediction results of the joint knowledge extraction model for each entity and relation in the test set are shown in the Table 6:

The calculation rules of the entity classification results shown in Table 6 are expressed as follows: the positive samples of the entities are the ones with the correct location information of the characters in the entity and the correct category information of the entity; the calculation rules of the relation classification results are expressed as follows: the positive samples of the relations are the ones with the correct location information of the characters in the entity and the correct relationship information of the entity. The reason for defining this calculation rule is that in the subsequent process

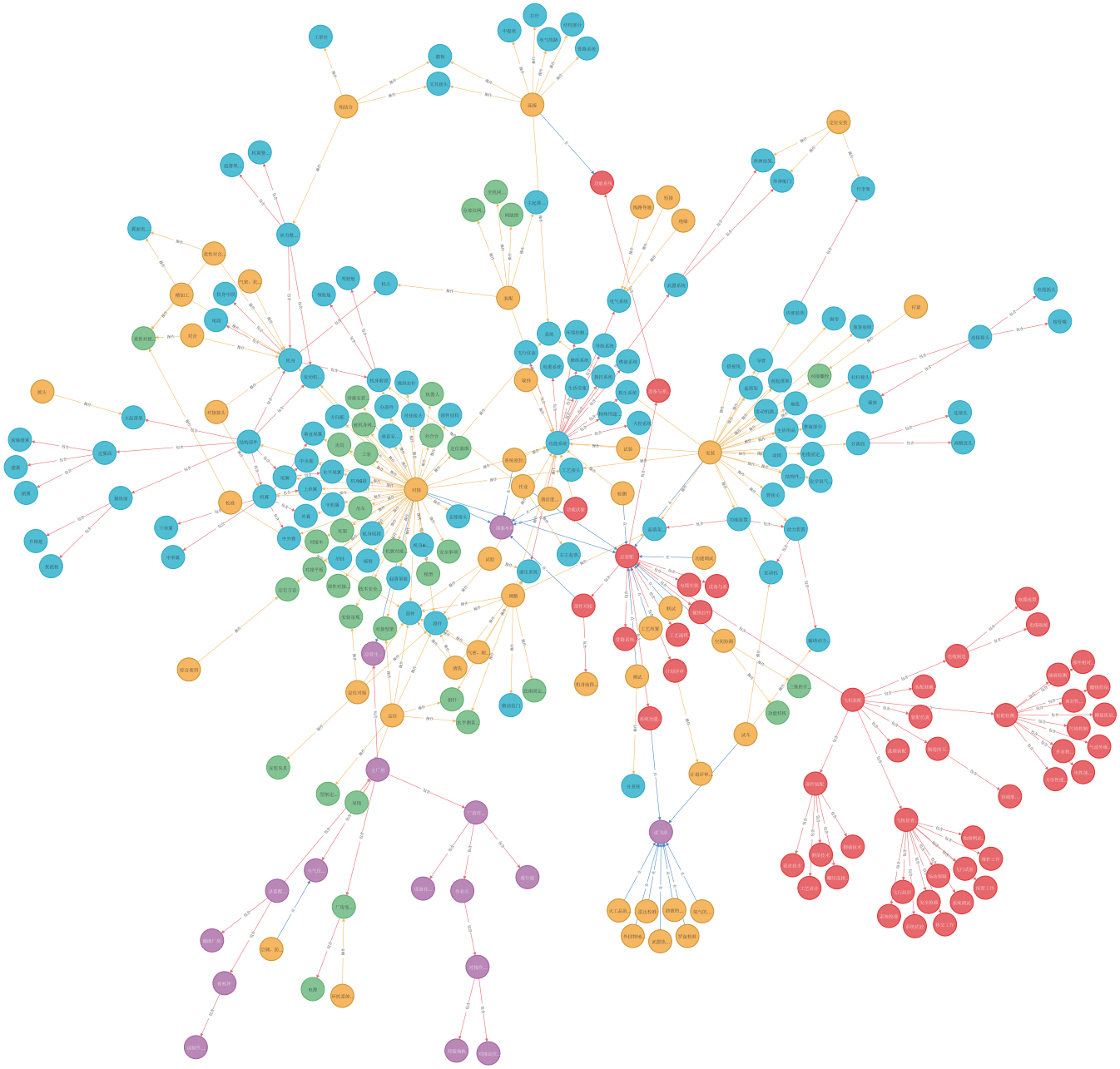

FIGURE 7. A knowledge graph example of aviation assembly.

of restoring the knowledge triples through the character annotations, the reconstruction scheme of relations and entities is relatively independent.

In the Table 6, the recognition performance of the joint knowledge extraction model for various entities and relations categories. By comparing the recognition performance of 5 kinds of entities and 3 kinds of relations, it can be found that the overall recognition effect is about 90%, therefor the effectiveness of the model can be proven.

Most of the recognition errors are caused by blurred entity boundaries, such as the "cleanliness detection" is an operation entity, but the "detection" is separately identified as an entity during the prediction process; the "component assembly workshop" is a facility entity, but the "component assembly" is identified as a step entity during recognition. It can be seen that the requirement of the knowledge extraction model in the professional domain is higher than the one in the general domain.

B. THE CONSTRUCTION OF THE KNOWLEDGE GRAPH IN THE AVIATION ASSEMBLY DOMAIN

KG are graph structures, where entities can be regarded as nodes and relations between the nodes can be regarded as edges [37]. The storage of knowledge graphs generally uses graph databases, and the extensive graph databases are Neo4j, Titans, OrientDB, etc.. In this paper, Neo4j database [38]

TABLE 6. Results statistics of various entities and relations on the AA corpus.

Entities or relations categories	Precision/%	Recall/%	F1-score/%
component	86.21	89.29	87.72
facility	100.00	100.00	100.00
operation	71.43	100.00	83.33
step	87.50	100.00	93.33
tool	100.00	66.66	80.00
entities average/total	89.75	89.68	89.71
Instrument-Agency	68.12	73.44	70.68
Component-Whole	100.00	62.07	76.60
Content-In	90.91	83.33	86.96
relations average/total	91.44	91.10	91.27

for the storage and presentation of the knowledge graphs are used. Neo4j is an open-source graph database that provides a friendly interface with python and supports various graph mining algorithms, providing a technical basis for knowledge graph applications.

In this paper, 1, 308 pairs of triples are formed from the entities and relations extracted by the joint knowledge extraction model. These triples are imported into the graph database via the Neo4j interface with python. Based on this data, we can construct the aviation assembly knowledge graph and display it on the visualization tool provided by Neo4j. Since the overall knowledge graph is too large, 300 nodes are selected as examples, which are shown in the Figure 7:

In the Figure 7, a knowledge graph consisting of five kinds of entities and three kinds of relations is presented. Here, the step entities are marked red, the operation entities are marked yellow, the facility entities are marked purple, the component entities are marked blue, and the tool entities are marked green. Similarly, relations are also distinguished by colors, where Component-Whole relations are marked red, Content-In relations are marked blue, and Instrument-Agency relations are marked yellow. Because the excerpt graph is still too large, the details of the assembly process cannot be directly reflected. Focus on a certain detail in the assembly process, such as aircraft structural components, the Figure 8 can be obtained.

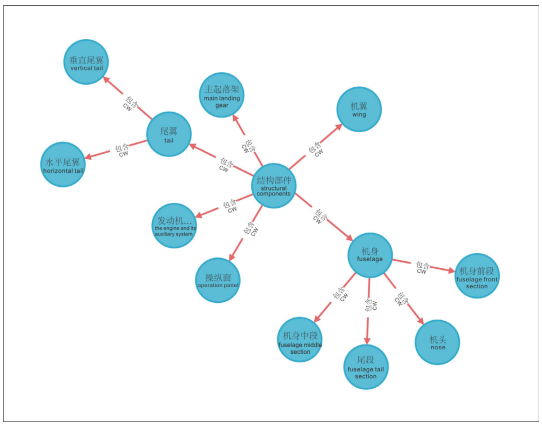

FIGURE 8. A sample knowledge graph of airplane structural components.

From the Figure 8, it is clearly that the relation between the aircraft structural components is described, further the corresponding real assembly components are shown in the Figure 9.

FIGURE 9. The assembly relation of the real aircraft structural components.

In the Figure 9, the real structural components such as the fuselage and the nose are present, and their relations CW are shown as well. Compared with the Figure 8, it is clear that the constructed KG is able to well describe the actual assembly scene.

V. CONCLUSION

In order to fully explore the intrinsic correlation value of massive data, and to comprehensively and accurately construct the KG in the aviation assembly domain is constructed from natural language texts. The process of KG automated generation is present in this paper.

Firstly, the text corpus and the architecture of entities and relations in the aviation assembly domain are innovatively proposed. With a novel tagging scheme, new data is present and rather reliable. Such complicated feature engineering in the case of the work consist in the first manifested contribution.

Then the paper proposes the joint knowledge extraction model with the approach of RL. The proposed joint knowledge extraction model is testified from the practical scenarios of the general assembly and component assembly. The experimental results show that the proposed model maintains high accuracy, recall, and F1-score without relying on manual features, with the F1-score of 89.71% for entities, the F1-score of 91.27% for relations, and the overall average F1-score of 82.41%. The model, which greatly improves the efficiency in construction tasks of professional domain knowledge graphs, has certain innovation in the construction of the professional domain corpus and the professional domain knowledge extraction task. The proposed model can complete the tasks at a high level, and continuously improve the efficiency of human-computer interaction. Thus the proposed model has a high practical application value.

Finally, in the paper, the knowledge graph, which included 1, 308 pairs of triples composed of five kinds of entities and three kinds of relations, is constructed in the aviation assembly domain. The graph describe the knowledge details of the assembly process well. Therefore it can be regard as the knowledge base and the technical basis of the corresponding assembly auxiliary system in the future.

In the future work, the fuzzy entity boundary, which is the main problem in the entity recognition process, is one of the relevant studies. The context consideration is the key to further improve the recognition effects. Moreover, this paper presents a basic and effective framework of entities and relations in the aviation assembly domain, and more complex classification strategies could be one of the future topics. For example, the consideration of particular or generic relations could be useful to improve the structure of the KG and the hierarchical task network could be implemented in this case.

REFERENCES

[1] A. Hogan, E. Blomqvist, and M. Cochez, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–37, 2021.
[2] M. He, X. Du, and B. Wang, "Representation learning of knowledge graphs via fine-grained relation description combinations," *IEEE Access*, vol. 7, pp. 26466–26473, 2019.
[3] R. Guha, D. Brickley, and S. Macbeth, "Schema.Org: Evolution of structured data on the web," *Commun. ACM*, vol. 59, no. 2, pp. 44–51, 2016.
[4] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Waltham, MA, USA: Elsevier, 2011.
[5] T. Al-Moslmi, M. G. Ocana, A. L. Opdahl, and C. Veres, "Named entity extraction for knowledge graphs: A literature overview," *IEEE Access*, vol. 8, pp. 32862–32881, 2020.
[6] X. Shi, X. Tian, J. Gu, F. Yang, L. Ma, Y. Chen, and T. Su, "Knowledge graph-based assembly resource knowledge reuse towards complex product assembly process," *Sustainability*, vol. 14, no. 23, p. 15541, Nov. 2022, doi: 10.3390/su142315541.
[7] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1411–1428, Oct. 2006.
[8] Y. Feng, H. Zhang, W. Hao, and G. Chen, "Joint extraction of entities and relations using reinforcement learning and deep learning," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–11, Aug. 2017.
[9] O. Felix, C. J. Paredis, and V.-H. Birgit, "Applying knowledge bases to make factories smarter," *Automatisierungstechnik*, vol. 67, no. 6, pp. 504–517, 2019.
[10] L. Yuan, Ed., *Aviation Manufacturing Engineering Manual: Aircraft Assembly*. Aviation Industry Press, 2010.
[11] J. Chen and C. Teng, "Joint entity and relation extraction model based on reinforcement learning," *J. Comput. Appl.*, vol. 39, no. 7, pp. 1918–1924, 2019, doi: 10.11772/j.issn.1001-9081.2019010182.
[12] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. Zhao, "Natural language question answering over RDF: A graph data driven approach," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2014, pp. 313–324.
[13] Z. Sun, G. U. Junzhong, and J. Yang, "Chinese entity relation extraction method based on deep learning," *Comput. Eng.*, vol. 44, no. 9, pp. 164–170, Jan. 2018.
[14] C. N. D. Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," *Comput. Sci.*, vol. 86, no. 86, pp. 132–137, 2015.
[15] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, and B. Xu, "Joint entity and relation extraction based on a hybrid neural network," *Neurocomputing*, vol. 257, pp. 59–66, Sep. 2017, doi: 10.1016/j.neucom.2016.12.075.
[16] Z. Geng, Y. Zhang, and Y. Han, "Joint entity and relation extraction model based on rich semantics," *Neurocomputing*, vol. 429, pp. 132–140, Mar. 2021, doi: 10.1016/j.neucom.2020.12.037.
[17] F. Li, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinf.*, vol. 18, no. 1, p. 198, Mar. 2017, doi: 10.1186/s12859-017-1609-9.
[18] Q. Li and H. Ji, "Incremental joint extraction of entity mentions and relations," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 402–412.
[19] M. Miwa and Y. Sasaki, "Modeling joint entity and relation extraction with table representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1858–1869.
[20] S. Singh, S. Riedel, B. Martin, J. Zheng, and A. McCallum, "Joint inference of entities, relations, and coreference," in *Proc. Workshop Automated Knowl. Base Construct.*, Oct. 2013, pp. 1–6.
[21] A. Katiyar and C. Cardie, "Going out on a limb: Joint extraction of entity mentions and relations without dependency trees," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 917–928.
[22] M. Miwa and M. Bansal, "End-to-end relation extraction using LSTMs on sequences and tree structures," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 567–574.
[23] S. Zheng, J. Xu, H. Bao, Z. Qi, Z. Jie, H. Hao, and X. Bo, "Joint learning of entity semantics and relation pattern for relation extraction," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2016, pp. 443–458.
[24] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder, "Joint entity recognition and relation extraction as a multi-head selection problem," *Exp. Syst. Appl.*, vol. 114, pp. 34–45, Dec. 2018.
[25] J. Feng, M. Huang, Z. Li, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Apr. 2019, pp. 1–8.
[26] P. Qin, W. Xu, and W. Y. Wang, "Robust distant supervision relation extraction via deep reinforcement learning," in *Proc. ACL*, Jul. 2018, pp. 2137–2147.
[27] R. Takanobu, T. Zhang, J. Liu, and M. Huang, "A hierarchical framework for relation extraction with reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7072–7079.
[28] G. A. Levow, "The third international Chinese language processing bakeoff: Word segmentation and named entity recognition," in *Proc. 5th SIGHAN Workshop Chin. Lang. Process.*, 2006, pp. 108–117.
[29] N. Peng and M. Dredze, "Named entity recognition for Chinese social media with jointly trained embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 548–554.
[30] Y. Zhang, Y. Wang, and J. Yang, "Lattice LSTM for Chinese sentence representation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1506–1519, 2020, doi: 10.1109/TASLP.2020.2991544.
[31] J. Chen and J. Gu, "Jointly extract entities and their relations from biomedical text," *IEEE Access*, vol. 7, pp. 162818–162827, 2019, doi: 10.1109/ACCESS.2019.2952154.
[32] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. (2018). *Doccano: Text Annotation Tool for Human*. [Online]. Available: https://github.com/doccano/doccano and https://github.com/doccano/doccano
[33] F. Camastra and A. Vinciarelli, *Markovian Models for Sequential Data*. London, U.K.: Springer, 2008, pp. 265–303, doi: 10.1007/978-1-84800-007-0_10.
[34] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.
[35] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*.
[36] L. Guo, F. Yan, T. Li, T. Yang, and Y. Lu, "An automatic method for constructing machining process knowledge base from knowledge graph," *Robot. Comput.-Integr. Manuf.*, vol. 73, Feb. 2022, Art. no. 102222, doi: 10.1016/j.rcim.2021.102222.
[37] X. Yan, F. Jian, and B. Sun, "SAKG-BERT: Enabling language representation with knowledge graphs for Chinese sentiment analysis," *IEEE Access*, vol. 9, pp. 101695–101701, 2021.
[38] Z. Wang, B. Zhang, and D. Gao, "A novel knowledge graph development for industry design: A case study on indirect coal liquefaction process," *Comput. Ind.*, vol. 139, Aug. 2022, Art. no. 103647, doi: 10.1016/j.compind.2022.103647.

PEIFENG LIU received the B.S. and M.S. degrees in automation and control engineering from the University of Duisburg–Essen, Duisburg, Germany, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in mechanical engineering with the State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology (HUST). His research interests include knowledge graph, intelligent manufacturing, and deep learning.

LU QIAN received the M.S. degree in mechanical engineering from Beijing Institute of Technology, Beijing, China, in 2014, and the Ph.D. degree in electrical engineering and information technology from the University of Duisburg–Essen, Duisburg, Germany, in 2019. She is currently a Lecturer with the School of Transportation and Logistics Engineering, Wuhan University of Technology (WHUT), Wuhan, China. Her research interests include fault diagnosis, process monitoring, and intelligent control.

XINGWEI ZHAO received the B.S. and M.S. degrees in mechanical engineering from the University of Duisburg–Essen, Duisburg, Germany, in 2012 and 2013, respectively, and the Ph.D. degree in mechanical engineering from the Technical University of Berlin, Berlin, Germany, in 2017. He is currently an Associate Researcher with the State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China. His research interests include nonlinear dynamics, robot control, and robotic manufacture.

BO TAO (Member, IEEE) received the B.S. and Ph.D. degrees in mechanical engineering from the Huazhong University of Science and Technology (HUST), in 1999 and 2007, respectively. From June 2007 to June 2009, he was a Postdoctoral Researcher. He was an Associate Professor and a Professor with the School of Mechanical Science and Engineering, HUST, in 2009 and 2013, respectively. From June 2013 to June 2014, he was a Visiting Scholar with the Department of Mechanical Engineering, University of California (UC) at Berkeley, Berkeley, CA, USA. He is currently a Changjiang Scholar Chair Professor with HUST. His research interests include intelligent manufacturing, robotics technologies, and the IoT technologies and applications.