US-China Education Review A, September 2025, Vol. 15, No. 9, 631-638
doi: 10.17265/2161-623X/2025.09.004

Construction of Vocational Education Knowledge Graph Based on DeepSeek

WANG Xiaoyu
Guangzhou Civil Aviation College, Guangzhou, China

In the context of the digital and intelligent era, exploring effective paths for the digital transformation of vocational education through artificial intelligence technology is of vital importance. This research takes the "Compressor" chapter of the "Civil Aircraft Maintenance Personnel License Management Rules" (CCAR-66-R3) license course "Gas Turbine Engine" as the object and proposes an automatic construction method for vocational education course knowledge graphs based on the DeepSeek large language model (LLM). By designing targeted prompt words (Prompts), the model API is invoked to extract (entity-relation-entity) triples from unstructured textbook texts, and Neo4j graph database is used for storage and visualization. Eventually, a structured knowledge network representing the structure, working principle, and fault diagnosis of the compressor system is formed. This method verifies the feasibility of large language models in the structured processing of complex technical field knowledge and provides a reference technical solution for the efficient construction of knowledge bases on the vocational education field and the support of intelligent teaching applications.

Keywords: knowledge graph, large language model, vocational education, triple extraction

Introduction

In the digital era, the rapid advancement of artificial intelligence technology has brought about new transformations in vocational education. The new round of industrial upgrading has set higher requirements for the knowledge structure and comprehensive qualities of technical and skilled talents. Against this backdrop, vocational education in China, as an important force supporting national strategies and regional economic development, its high-quality development has become a key link in the national modernization process. Promoting the digital transformation of vocational education and building a smart education ecosystem that meets the needs of the new era are the key directions for improving the quality of talent cultivation and enhancing the ability to serve industries.

However, the vocational education curriculum system, especially in high-end and complex technical fields such as aviation maintenance and intelligent manufacturing, generally exhibits characteristics such as a complex knowledge system, abstract concepts, and high skill requirements (Liu, Zhou, & Li, 2025). The traditional course organization method is unable to intuitively reveal the internal logic and hierarchical structure of knowledge points, resulting in problems such as low efficiency in knowledge transmission and heavy cognitive load for

Acknowledgement: Special Project of the Scientific Research Backbone Training Program of Guangzhou Civil Aviation College, "Research on the Construction of Course Knowledge Graphs Under the Background of Educational Digital Transformation" (Project number: 24X4134).
WANG Xiaoyu, Lecturer, Guangzhou Civil Aviation College, Guangzhou, China.

students during the teaching process (Gao & Mu, 2024). Knowledge graph, as a tool for structuring and managing knowledge, through the triad form of "entity—relationship—entity", can integrate fragmented knowledge into an interrelated semantic network, thereby providing core underlying support for the construction of intelligent educational applications (Yang & Rai, 2025). Therefore, constructing a professional knowledge graph in the vocational education field is of great value in solving the teaching problems of aviation maintenance and achieving precise push of course learning.

Although knowledge graphs have significant advantages, their construction process is rather cumbersome. Traditional methods rely on domain experts to manually build them, which has bottlenecks such as high labor costs, long cycles, and poor scalability (Xiao, Wang, Guo, & Luo, 2022). In recent years, artificial intelligence technologies represented by large language models (LLMs) have made breakthrough progress, demonstrating extraordinary capabilities in natural language understanding, information extraction, and text generation, providing a new technical paradigm for the automated and large-scale construction of domain knowledge graphs (Liu & Hao, 2024). Large language models can accurately extract entities and their relationships from unstructured professional texts, significantly improving the construction efficiency of knowledge graphs and making their application in specific vertical fields possible.

This research takes the training materials of the "Civil Aircraft Maintenance Personnel License Management Rules" (CCAR-66-R3) issued by the Civil Aviation Administration of China (CAAC) as the research object. This license is the basic qualification for aviation maintenance personnel to enter the workplace, and its designated training materials are rigorous in content and complex in structure, making it a typical knowledge-intensive course in vocational education. DeepSeek, as a high-performing domestic large language model, possesses strong Chinese semantic understanding and reasoning capabilities (Dai, Wu, & Zhan, 2025). Exploring the use of the DeepSeek model to automatically process the content of the license training materials and construct a knowledge graph for the maintenance license course is not only an important practice for empowering vocational education teaching reform with large models, but also can provide a reference for the construction of knowledge graphs for similar courses.

Literature Review

With the continuous deepening of the integration of artificial intelligence and education, knowledge graphs, as a core technology for structured knowledge representation and intelligent applications, have gradually evolved into a focal area of research within the field of education. Especially in vocational education, knowledge graphs can effectively support the structured organization of course knowledge, personalized learning path recommendation, and intelligent teaching assistance, and have significant theoretical and practical value.

In terms of the theoretical basis of educational knowledge graphs, Li Zhen and Zhou Dongdai (2019) pointed out that knowledge graphs are an important development of the symbolic paradigm in the era of big data and artificial intelligence. The core lies in constructing semantic networks through "entity-relation-entity" triples to model and manage educational knowledge. Educational knowledge graphs can be divided into static knowledge graphs (SKG) and dynamic reasoning graphs (DRG). The former focuses on the structured representation of subject knowledge, while the latter pays attention to the dynamic cognitive and behavioral relationships in the teaching process. This research further proposed a multi-dimensional perspective of educational knowledge graphs, including knowledge modeling, resource management, knowledge navigation, and learning cognition, providing a theoretical framework for subsequent research.

At the technical implementation level, the construction methods of knowledge graphs mainly include top-down and bottom-up approaches, involving key steps such as knowledge extraction, integration, storage, and reasoning (Zhang et al., 2024; Deng, 2022; Hang, Feng, & Lu, 2021). Huang Huan, Yuan Shuai, He Ting-ting, and Wu Linjing (2019) adopted a combination of manual and semi-automatic methods when building the knowledge graph for the "Fundamentals of Java Programming" course. They defined three-level knowledge units of "chapter—section—knowledge point" and their semantic relationships (including, sequence, and correlation), and implemented ontology modeling based on the Protege tool. Xie Jun, Yang Haiyang, Liang Fengmei, and Xu Xinying (2023), on the other hand, in the "Signals and Systems" course, defined three types of entities: "knowledge point—example—question", and combined word and sentence similarity algorithms to develop an intelligent Q&A system based on WeChat mini-programs, significantly enhancing students' learning autonomy and Q&A efficiency.

In terms of application practice, Zhang Huinan (2023) took "Principles of Computer Organization" as an example and constructed a three-dimensional ontology model integrating course objectives, knowledge points, and resources. The BiLSTM-CRF model was adopted for entity recognition, and the knowledge storage and visual query were ultimately realized through the Neo4j graph database. This research also verified the positive effects of the system in enhancing students' learning motivation and knowledge acquisition efficiency through questionnaires and teacher interviews.

However, the existing research still has certain limitations: Most of the achievements are concentrated on general education or higher education professional courses, while there are relatively few studies on vocational education, especially vocational qualification certification courses (such as the CCAR-R3 license course); the construction of knowledge graphs mostly relies on manual annotation and expert knowledge, and the degree of automation still needs to be improved; there is a lack of systematic methodologies and practical verification in dynamic updates, multi-modal knowledge fusion, and quantitative assessment of learning effects.

Research Method

Data Selection

The data for this research are sourced from Module M5, "Gas Turbine Engine", in the training materials designated by the Civil Aviation Administration of China (CAAC) in accordance with CCAR-66-R3, the "Regulations for the Management of Civil Aircraft Maintenance Personnel Licenses". The selection of this material as the data source is primarily based on three points: (1) Textual authority: This textbook was compiled by the Civil Aviation Maintenance Association of China (CAMAC) and approved by the Flight Standards Department of the Civil Aviation Administration. It is the officially designated textbook for the civil aircraft maintenance personnel license examination, and its content is highly authoritative and industry-standardized, ensuring the accuracy and reliability of the research data. The knowledge graph constructed based on this ensures the accuracy and authority of the knowledge, laying a trustworthy foundation for its subsequent application in teaching and practice. (2) Clear structure: "Gas Turbine Engine" has a rigorous knowledge system structure, abstract concepts, complex component interrelations, and contains a large amount of working principles, typical faults, and maintenance norms. (3) Knowledge intensity: The knowledge content includes a large number of entity concepts and various complex relationships, highly representing the typical characteristics of complex knowledge, strong logic, and high safety requirements in aviation maintenance vocational education. It is an ideal sample for verifying the ability of large language models to handle professional field texts. Based on this, the fourth chapter, "Compressor", in "Gas Turbine Engine" was ultimately selected as the typical sample text for this research.

Data Processing

To ensure the quality of the data input into the large language model and improve the accuracy of knowledge extraction, it is necessary to preprocess the original textbook content. Since the original textbook is in PDF format, a professional PDF parsing library (such as PyMuPDF or pdfplumber) is first used to extract the text. Compared with conventional OCR recognition, this method can better preserve the structure and sequence of the text and minimize character errors caused by format disorder (Hu, 2022). Then, based on the directory structure of the textbook, the start and end page numbers of the "Compressor" chapter are located, and the corresponding text is completely extracted to form an independent text library as the original data for subsequent processing. Additionally, noise information such as headers, footers, page numbers, chart titles (retaining them as text descriptions), and pure numerical numbering that is irrelevant to the core knowledge content need to be removed. Full-width characters (such as Chinese colons and parentheses) are converted to half-width characters to ensure the consistency of the text format. The expressions of key terms in the textbook are manually checked and unified.

Large language models have restrictions on input length, and overly long texts can affect their understanding and extraction accuracy. A "semantic boundary-based" segmentation strategy (Wang, Tang, & He, 2014) can be adopted. This strategy does not simply cut text by a fixed number of characters but takes natural paragraphs as the core unit and appropriately combines full stops (.) as boundary points to split long paragraphs into short ones with complete semantic meanings. The length of each text paragraph is controlled between 200 and 500 characters to ensure that each paragraph describes a relatively complete knowledge scenario (such as defining a concept, describing the structure of a component, or explaining a working principle), thereby providing the model with the best context information.

Knowledge Extraction Based on DeepSeek API

Knowledge extraction is the core link in building a knowledge graph. This research adopts the DeepSeek large language model as the core engine and uses its provided API interface to extract structured triples from preprocessed text. The capabilities of large language models are highly dependent on the input prompt words (Prompt) (R. Y. Cao, & S. J. Cao, 2023). To ensure the accuracy and standardization of the extraction, after multiple rounds of debugging and optimization, the final Prompt template is determined as shown in Table 1.

Table 1

Prompt Words for Triple Extraction in Knowledge Graphs

你是一个资深的航空发动机专家，也是知识图谱工程师。你的任务是从以下关于燃气涡轮发动机压气机的技术文本中，精确抽取出所有有价值的三元组知识。 You are a seasoned expert in aviation engines and also a knowledge graph engineer. Your task is to precisely extract all valuable triple knowledge from the following technical text about gas turbine engine compressors. **extraction rules:** 1. **entity identification:** Identify key entities in the text, such as components (e.g., rotor blades, intake casing), states (e.g., surge, stall), performance parameters (e.g., boost ratio, efficiency), etc. 2. **defining relations:** Determine the relationship between entities based on the context. The relationship types are limited to a predefined set, such as: "is a component of", "has the function of", "will lead to", "is located in", "is divided into", "is characterized by", "is used to prevent", etc. 3. **output format:** Output strictly in JSON list format. Each element in the list is a dictionary, containing and only containing three keys: "subject", "relation", "object". Make sure the output is only in JSON format and without any explanatory text. **text to be processed:** {insert_text_segment} Please start the extraction:

Automate batch processing through Python by calling API. Write a Python script to sequentially fill the list of preprocessed text paragraphs into the {insert_text_segment} position in the above Prompt template. Use the requests library to loop through the DeepSeek API calls and set reasonable request parameters (temperature = 0.1 to ensure the determinacy of the output, max_tokens to control the generation length). For each API response, parse it using json.loads() and extract the list of triples.

Knowledge Graph Construction and Storage

To construct a queryable and computable knowledge graph from the discrete triples extracted from the text, two steps are required: schema design and storage.

First, the schema pattern design is carried out. The pattern is the abstract framework of the knowledge graph, defining the entity types and the relationship types among them. Based on the CCAR-R3 textbook and the knowledge in the field of aero engines, the ontology pattern was designed, mainly defining the following core entity types and relationship types:

Entity type: 部件 (Component), 系统 (System), 故障 (Fault), 性能参数 (Parameter), 操作(Operation).

Relationship type: as Part (包含部件), is A (属于/是某种类型), leads To (导致), has Function (具有功能), has Characteristic (具有特性), prevent By (通过预防).

Secondly, a graph database is chosen for storage and visualization. In this study, Neo4j is selected as the storage and visualization tool for the knowledge graph. Neo4j is a native graph database, and its property graph model is highly intuitive, which is highly compatible with our triple data model and supports the efficient graph traversal query language Cypher. The cleaned triple data are batch-imported into the Neo4j database through the py2neo library of Python. In this process, the subjects and objects in the triples are created as graph nodes (Node) and labeled with corresponding entity type tags; the relations (Relation) are created as directed edges connecting the nodes and are attached with relation type attributes. After the import, the built-in visualization tool of Neo4j can be used to visually display the knowledge network structure of the “Compressor” chapter, preliminarily verifying the correctness and richness of the graph construction.

Knowledge Graph Construction Based on DeepSeek

Triple Extraction

By using the triplet extraction prompt words designed in the previous text, the DeepSeek API is called in Python to extract the subject-predicate-object triplets from the chapter “Compressor” in M5 “Gas Turbine Engine”, which serves as the basic framework for constructing the knowledge graph. Some of the extraction results are shown in Table 2.

Table 2

Compressor Triplets (Partial Data)

Subject	Relation	Object	Subject	Relation	Object
压气机 compressor	具有功能 have ... function	对空气进行增压 pressurize the air	轴流式压气机 axial flow compressor	具有性能参数 have performance parameters	增压比 pressure ratio
压气机 compressor	分为 fall into	轴流式压气机 axial flow compressor	轴流式压气机 axial flow compressor	可能导致 lead to	喘振 surge
压气机 compressor	分为 fall into	离心式压气机 centrifugal-flow compressor	喘振 surge	具有症状 have symptoms	发动机振动加大 the engine vibration has increased

Construction of Knowledge Graph

This research utilized the py2neo library of Python to import the obtained triple data into the Neo4j graph database, completing the construction and storage of the knowledge graph. By invoking the API of large language models through the Python language, the construction of the knowledge graph for license textbooks was achieved. Some prompt words and codes for the construction of the knowledge graph are shown in Table 3 as follows.

Table 3

Knowledge Graph Construction Prompt Words and Codes (Part)

```
from py2neo import Graph, Node, Relationship
# 连接Neo4j图数据库
graph = Graph (“bolt://localhost:7687”, auth = (“neo4j”, “your_password”))
# 定义知识图谱模式（实体标签与关系类型）
labels = {“部件”, “故障”, “性能参数”, “操作”} # 部分实体类型示例
relation_types = {“是组成部分”, “具有功能”, “会导致”, “分为”} # 部分关系类型示例
# 构建知识图谱的主函数
def build_knowledge_graph (triples):
    tx = graph.begin ()
    nodes_dict = {} # 用于缓存已创建的节点，避免重复创建
    for triple in triples:
        # 创建或获取主语节点
        subj_node = nodes_dict.get (triple[“subject”])
        if not subj_node:
            subj_node = Node (labels, name = triple [“subject”])
            nodes_dict [triple [“subject”]] = subj_node
            tx.create (subj_node)
        # 创建或获取宾语节点
        obj_node = nodes_dict.get (triple [“object”])
        if not obj_node:
            obj_node = Node (labels, name = triple [“object”])
            nodes_dict [triple [“object”]] = obj_node
            tx.create (obj_node)
        # 创建关系
        rel = Relationship (subj_node, triple [“relation”], obj_node)
        tx.create (rel)
    tx.commit () # 提交事务，将所有数据一次性写入图数据库
# 从CSV文件读取清洗后的三元组数据
import pandas as pd
triples_df = pd.read_csv (“compressor_knowledge_triples.csv”)
triples = triples_df.to_dict (“records”)
# 执行图谱构建
build_knowledge_graph (triples)
print (“知识图谱构建完成”)
```

Based on the above triples, a simplified knowledge graph can be drawn by calling the DeepSeek API, showing the core entities and their relationships. Since the complete graph is very complex, here we mainly display the “compressor” and its directly related core concepts, as shown in Figure 1.

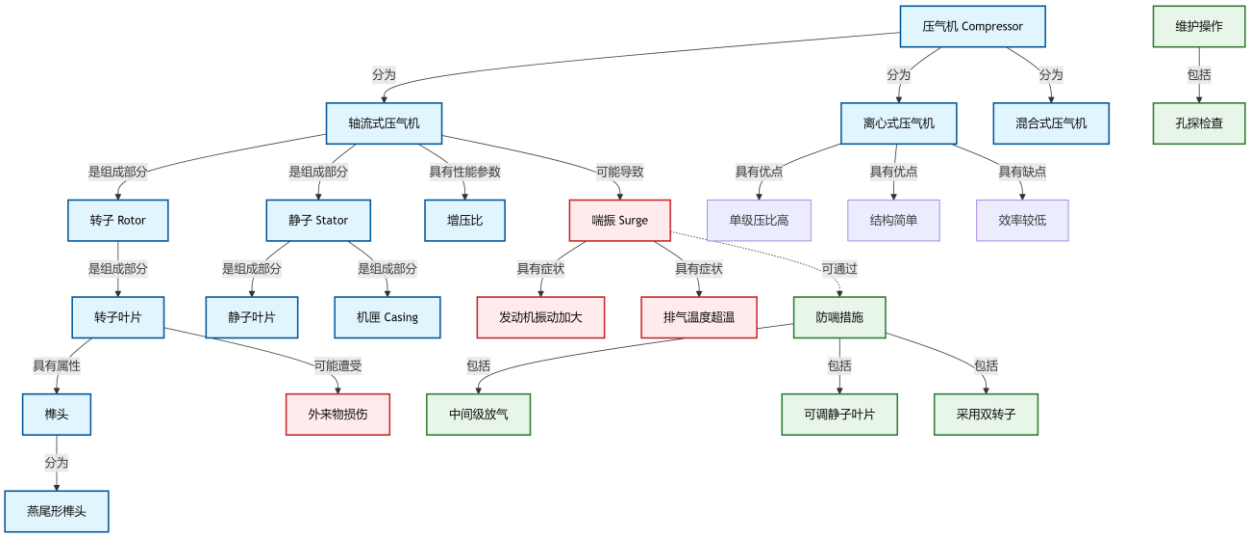

Figure 1. Compressor knowledge graph (partial data).

Conclusion and Prospect

Main Conclusion

This research selected the DeepSeek large language model to address the issue of constructing knowledge graphs for vocational education courses. Taking the “Compressor” chapter from the CCAR-66R3 license course “Gas Turbine Engine” as the research object, it achieved the transformation from unstructured textbook text to structured knowledge graphs. Through prompt design, the model can effectively identify core entities such as “rotor”, “stator”, and “surge”, as well as semantic relationships such as “is a component of”, “will cause”, and “has the function of” from professional textbooks, forming an initial structured knowledge network. Compared with the manual construction by experts, the use of large language models can significantly reduce the initial human and time costs of knowledge graph construction, providing a solution for the rapid construction of large-scale knowledge bases in the vocational education field. This research has formed a complete technical path including data preprocessing, prompt design, API invocation, data cleaning, and graph storage. This technical process is not limited to the aviation maintenance field and can be transferred to other technical and skill fields with appropriate adjustments.

Limitations and Prospects

The knowledge graph constructed in this study has not yet been systematically validated for accuracy, completeness, and logical consistency by domain experts. Issues such as “hallucinations” and incorrect relation extraction that large language models may generate still need to be quantitatively evaluated and corrected through expert manual assessment. The current knowledge graph mainly includes explicit knowledge explicitly stated in textbooks, but lacks sufficient mining of implicit knowledge that requires deep reasoning. It has not yet achieved complex knowledge reasoning functions. In the future, a group of senior teachers and engineers specializing in aviation engines can be invited to form an expert panel to conduct a systematic reliability and validity test on the knowledge graph constructed in this paper. By calculating quantitative indicators such as precision, recall, and F1 value, and comparing them with the results of manual annotation, the performance boundaries and optimization directions of this method can be scientifically evaluated. It is possible to attempt to fine-tune the

DeepSeek model for domain adaptability using professional corpora in the aviation maintenance field to further improve its accuracy and reliability in extracting specific professional terms and complex relations.

References

Cao, R. Y., & Cao, S. J. (2023). The effect and enlightenment of ChatGPT to complete the task of knowledge organization. *Information and Documentation Services, 44*(5), 18-27.

Dai, G. H., Wu, X. G., & Zhan, W. H. (2025). Research on the solution and development of DeepSeek local deployment in terminals. *Mobile Communications, 49*(3), 100-106.

Deng, Z. J. (2022). AI-based knowledge graph construction technology and its application. *Radio Engineering, 52*(5), 766-774.

Gao, C. J., & Mu, S. (2024). Virtual simulation teaching strategy to promote deep learning of vocational skills—Taking "aircraft maintenance professional teaching" as an example. *Laboratory Research and Exploration, 43*(12), 165-171.

Hang, T.-T., Feng, J., & Lu, J.-M. (2021). Knowledge graph construction techniques: Taxonomy, survey and future directions. *Computer Science, 48*(2), 175-189.

Hu, T. (2022). Design and implementation of automatic compilation of maintenance work card based on Python. *Aviation Maintenance & Engineering, 67*(6), 59-62.

Huang, H., Yuan, S., He, T.-T., & Wu, L. J. (2019). Research on the construction of course knowledge graph for adaptive learning system—Taking "Java Programming Foundation" course as an example. *Modern Educational Technology, 29*(12), 89-95.

Li, Z., & Zhou, D. D. (2019). Research on conceptual model and construction method of educational knowledge graph. *e-Education Research, 40*(8), 78-86+113.

Liu, C., Zhou, K., & Li, J. L. (2025). Development and practice of new form textbooks for aircraft maintenance specialty under the background of integration of production and education. *Guangdong Vocational Technical Education and Research, 16*(4), 96-99+143.

Liu, S. N. Y., & Hao, X. H. (2024). The challenges and approaches of generative artificial intelligence in promoting educational innovation. *Tsinghua Journal of Education, 45*(3), 1-12.

Wang, M., Tang, X. L., & He, T. T. (2014). Research on a multi-document summarization method based on text segmentation technology. *Computer Applications and Software, 31*(9), 40-44.

Xiao, M. S., Wang, M., Guo, Y. Q., & Luo, J. M. (2022). Research on the construction of knowledge graph of programming language course. *Journal of Gannan Normal University, 43*(6), 95-100.

Xie, J., Yang, H. Y., Liang, F. M., & Xu, X. Y. (2025). Design and development on intelligent question & answering system based on the course graph—Taking signals and systems as an example. *Chinese Journal of Systems Science, 33*(3), 161-167.

Yang, Y., & Rai, S.-N. (2025). Research on knowledge graph based on collaborative knowledge construction conversation. *Modern Educational Technology, 35*(8), 97-106.

Zhang, H. N. (2023). Construction and application of curriculum knowledge graph for intelligent education (Chinese Master's theses full-text database, 2023).

Zhang, X. S., Liu, L., Wang, H. L., Su, G. B., & Liu, J. (2024). Survey of entity relationship extraction methods in knowledge graphs. *Journal of Frontiers of Computer Science and Technology, 18*(3), 574-596.