Conversational Knowledge Extraction from Technical Manuals: An LLM-based Framework with Ontological Guidance

Celeste Sguera[1], Antonio Pellicani[1,3], Giuseppe Modugno[4], Angelo Giannoccaro[4], and Michelangelo Ceci[1,2,3]

[1] Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy
[2] Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[3] Big Data Lab, National Interuniversity Consortium for Informatics, Rome, Italy
[4] MTM Project Srl, Monopoli (BA), Italy
c.sguera2@phd.uniba.it, {antonio.pellicani, michelangelo.ceci}@uniba.it, {modugno, giannoccaro}@mtmproject.com

Abstract. The effective extraction and structuring of knowledge from technical documentation remains a significant challenge in both educational and industrial contexts. Traditional information extraction approaches show limitations in capturing complex relationships and domain-specific terminology present in procedural manuals. This paper presents a novel intelligent framework that combines Large Language Models with ontological guidance for automatic extraction of entities and semantic relationships from educational manuals. The system integrates preprocessing and indexing of manuals, knowledge extraction based on Retrieval-Augmented Generation with ontological constraints, and conversational interaction for real-time procedural guidance. Evaluation across ten diverse scientific manuals demonstrated strong performance in tasklist extraction, question-answering, and action validation tasks.

Keywords: Large Language Models · Retrieval Augmented Generation · Ontologies.

1 Introduction

Over the past few decades, technology-enhanced learning environments have undergone a profound transformation, driven by the rapid advancement and widespread integration of digital technologies that have fundamentally reshaped the educational landscape. From the early stages of computer-assisted instruction in the mid-20th century to the current era of immersive learning experiences enabled by virtual and augmented reality, technology has assumed a central role in enhancing educational practices at all levels [2]. In addition, the seamless integration of multimedia resources, interactive simulations, and sophisticated online collaboration tools has fostered the development of engaging, flexible, and personalized learning environments tailored to the diverse needs of learners.

A prominent example of this pedagogical shift is the widespread adoption of interactive whiteboards and immersive educational games, which have contributed to the democratization of educational access and enabled students to participate more actively in the learning process, thereby developing essential skills such as creative thinking and problem-solving [1].

The transformative use of technology in education finds a parallel in the context of Industry 4.0, where managing technical knowledge is critical for automation, maintenance, and workforce training. In industrial environments, identifying domain-specific entities within manufacturing processes enables professionals to apply contextual knowledge effectively, facilitates the onboarding of new employees, and supports informed decision-making [6]. A similar approach is increasingly adopted in educational settings, where technologies such as augmented reality, virtual reality, and generative artificial intelligence are used to support both conceptual learning and the practical execution of experiments [2].

Despite these technological advances, the effective extraction and structuring of knowledge from vast repositories of unstructured technical documentation remains a significant challenge in both educational and industrial contexts [7]. Manufacturing manuals, maintenance guides, and instructional materials contain rich semantic information about entities, relationships, and procedures, yet this knowledge often remains locked in natural language formats that are difficult to process systematically. Traditional information extraction approaches, such as Named Entity Recognition (NER), have shown limitations in capturing the complex relationships and domain-specific terminology present in technical documentation, creating a need for more sophisticated approaches that can leverage both the contextual understanding capabilities of Large Language Models (LLMs) and the structured knowledge representation provided by domain ontologies [5]. This limitation not only hampers efficient knowledge transfer and procedural training but also creates barriers to automated quality assurance and intelligent tutoring systems that guide users through technical procedures [3].

In this paper, we present a novel intelligent framework that combines LLMs with ontological guidance for the automatic extraction of entities and semantic relationships from educational manuals, aimed at assisting users in the guided execution of technical procedures, such as scientific experiments or manufacturing processes. The proposed methodology is structured into three interconnected core modules: *i)* a preprocessing and semantic indexing pipeline that transforms unstructured technical documentation into structured representations stored in a vector database for efficient retrieval; *ii)* a knowledge extraction module based on Retrieval-Augmented Generation (RAG), where domain-specific ontologies guide and validate the extraction process to ensure semantic consistency and completeness; *iii)* a conversational interaction component that provides real-time user assistance through dynamic communication with the LLM, enabling guided execution of technical procedures.

The rest of the paper is organized as follows: Section 2 presents the proposed framework in detail, describing the three core modules and their integration; Section 3 evaluates the performance of the framework on three different knowledge

extraction tasks across scientific manuals; finally Section 4 draws conclusions and outlines future research directions.

2 Methodology

The proposed framework consists of an intelligent system based on LLMs for the automatic extraction of entities and semantic relationships from educational manuals, aiming to support users in the guided execution of technical procedures, such as scientific experiments or manufacturing protocols.

Our approach addresses the challenge of transforming unstructured technical documentation into actionable, structured knowledge that can dynamically guide users through complex procedures. The framework leverages the semantic understanding capabilities of LLMs while incorporating domain-specific ontological constraints to ensure extraction accuracy and consistency. Figure 1 illustrates the overall architecture of our system. The workflow begins with documentation preprocessing and semantic indexing to create structured representations. Users then interact with the conversational component, which queries a vector database to retrieve contextually relevant information and provides it to the LLM for generating guided instructions and feedback. This human-in-the-loop design enables dynamic, adaptive support for executing complex procedural tasks.

2.1 Preprocessing and Semantic Indexing Pipeline

The preprocessing pipeline transforms unstructured PDF manuals into semantically searchable knowledge representations through a multi-stage process designed to preserve contextual information while optimizing retrieval efficiency.

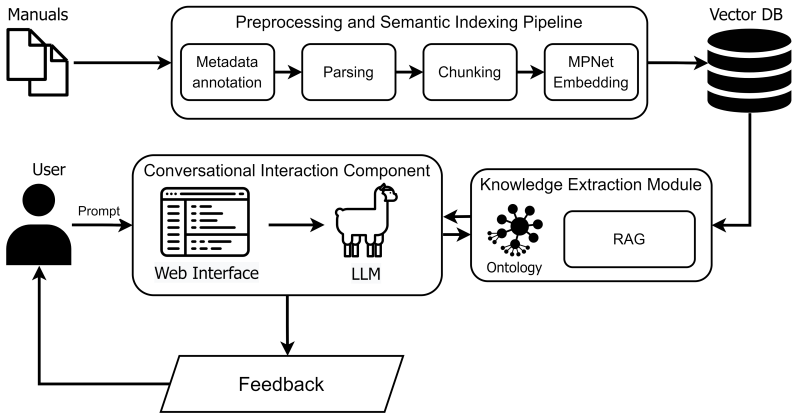

Fig. 1. Architecture of the LLM-based knowledge extraction system with ontological guidance. The system comprises three interconnected modules supporting real-time procedural guidance from technical documentation.

Technical and educational manuals are processed using the *pdfplumber*[1] library for accurate page-by-page content extraction. Each extracted text section is annotated with positional metadata using headers of the format [[PAGE X]] to maintain precise semantic references to the original document structure, ensuring that retrieved information can be traced back to its source location.

The extracted text undergoes parsing to isolate procedurally relevant information, including step-by-step instructions, objectives, required tools, and operational guidelines. This filtering process removes extraneous material while preserving essential procedural knowledge for optimal language model processing. Then, the resulting structured content, enriched with metadata and unique identifiers, is segmented into 1000-character chunks with 100-character overlaps to maintain contextual continuity across boundaries.

Finally, each text chunk is transformed into dense vector representations using the MPNet model [8], selected for its proven effectiveness with technical and scientific content [9]. The resulting embeddings are indexed in a FAISS vector database[2], enabling efficient similarity-based retrieval and providing the foundation for contextual query processing in the subsequent interaction component.

2.2 Knowledge Extraction Module

The knowledge extraction module operates through a RAG [4] framework enhanced with domain-specific ontological guidance to ensure accurate entity identification and relationship extraction from technical documentation.

When a user requests information about a specific procedure or concept, the RAG mechanism performs similarity-based searches over the semantically indexed document chunks. The system retrieves the most relevant document segments from the FAISS vector database based on semantic similarity, which then serve as contextual input to guide the language model during response generation. This retrieval-augmented approach significantly reduces hallucinated outputs and enhances factual accuracy in knowledge-intensive tasks.

To ensure semantic consistency and domain-specific accuracy, the system additionally integrates user-defined ontologies that define conceptual taxonomies specific to the technical domain. These ontologies establish hierarchical relationships between entities such as instruments, materials, procedures, and operational parameters, providing structured constraints for the extraction process.

During extraction, the language model operates under dual guidance: leveraging both the retrieved contextual information from relevant documents and the ontological constraints from the domain taxonomy. This ensures that identified entities and their relationships maintain consistency with the predefined domain structure while remaining grounded in the source documentation.

The ontological guidance operates at multiple levels: first, by constraining the types of entities that can be extracted based on the domain taxonomy; second, by ensuring that identified relationships conform to the ontological structure; and

[1] https://github.com/jsvine/pdfplumber
[2] https://github.com/facebookresearch/faiss

third, by providing semantic validation for extracted knowledge. This approach is particularly valuable in technical domains where precision and consistency are critical, as it introduces an additional semantic layer that enhances disambiguation and classification accuracy.

2.3 Conversational Interaction Component

The conversational interaction component serves as the primary interface between users and the knowledge extraction system, facilitating procedural guidance in real-time through natural language communication. It processes user queries, coordinates with the knowledge extraction module, and delivers contextually relevant responses to support the guided execution of technical procedures.

The interaction workflow begins when users submit queries related to specific procedural steps, tool requirements, or troubleshooting scenarios. These queries are processed through the system's API interface, which maintains session context to enable coherent multi-turn conversations with the LLM. The component distinguishes between two primary interaction modes: informational queries, where users seek specific knowledge from the technical documentation, and procedural guidance requests, where users require step-by-step assistance during the execution of a specific task.

Upon receiving a user query, the component invokes the knowledge extraction module to retrieve relevant information from the preprocessed documentation. This contextual information, enhanced by ontological constraints, is then fed to the LLM, which generates domain-specific responses based on both the retrieved content and its inherent language understanding capabilities. The LLM maintains conversation history and procedural context, enabling it to provide sequential help that adapts to the user's position within a complex procedure.

It is worth noting that the conversational interaction component ensures that provided guidance aligns with the source documentation and ontological structure, minimizing the risk of procedural errors while maintaining the flexibility necessary for dynamic user interaction. The interface supports both synchronous question-answering and asynchronous procedural monitoring, where users can report completed actions and receive confirmation or corrective guidance. Through this human-in-the-loop design, the conversational component transforms static technical documentation into an interactive guidance system, enabling users to navigate complex procedures with confidence while maintaining traceability to authoritative source materials.

3 Experimental Evaluation

The evaluation was conducted using LLaMA 3.1 8B as the core language model. The dataset consists of ten technical manuals describing complete experimental procedures from various scientific domains, including life sciences, thermodynamics, biology, electromagnetism, and mechanics, provided by MTM Project srl.

These manuals were selected to represent diverse technical vocabularies, procedural complexities, and domain-specific terminologies. Each document contains information about experimental objectives, required equipment, procedural rules, execution phases, and step-by-step instructions. To support the ontological guidance component, domain-specific ontologies were automatically generated for each manual using GPT-4, which analyzed the technical content and created structured taxonomies of entities, relationships, and procedural concepts specific to each scientific domain.

The system's performance was assessed through three distinct knowledge extraction tasks that progressively test different aspects of procedural understanding: ***i)*** **tasklist extraction** requires the system to identify and provide all procedural steps necessary to perform a specific scientific experiment, ***ii)*** **question-type prompts** evaluate the ability of the system to respond to general informational queries about experiments, while ***iii)*** **action-type prompts** test its capacity to provide procedural guidance and validate completed actions. More specifically, for tasklist extraction, five prompts per document were classified as *correct* (complete match with expected tasklist), *partial* (mostly correct with minor omissions or inaccuracies), or *incorrect* (significant deviation from expected results). Semantic understanding was evaluated through standardized question-type prompts, specifically: *"List the phases of the experiment"* and *"What is the objective of the experiment?"*, with each question posed five times per document to assess consistency and accuracy. Finally, action-type prompts employed a validation approach using ten prompts per document, equally split between correct and incorrect procedural statements, to test the model's discriminative capabilities. Across all 250 evaluations, domain experts assessed responses according to technical accuracy standards.

The experimental evaluation demonstrates strong performance across all three knowledge extraction tasks, as shown in Table 1. The system achieved the highest accuracy in **tasklist extraction** with an overall success rate of 94%, with 72% fully correct responses and an additional 22% partial matches. **Question-type prompts** followed closely with 92%, while **action-type prompts** showed 91% correct responses.

Tasklist extraction was the task that showed more varied results, with the combined acceptable response rate of 94% (correct + partial) demonstrating that the system successfully captures most procedural steps, though complete accuracy remains challenging for complex multi-step procedures. Only 6% of responses were classified as incorrect, indicating minimal risk of providing misleading procedural information. Challenges were more frequent for more com-

Table 1. Expert evaluation results across all knowledge extraction tasks

Task Type	Results	Accuracy
Tasklist Extraction	36/50 correct and 11/50 partial	94%
Question-type Prompts	92/100 correct	92%
Action-type Prompts	91/100 correct	91%

Table 2. Detailed results for Question-type and Action-type Prompts

Task Type	Prompt Type	Results	Accuracy
Question-type Prompts	List experiment phases	44/50	88%
	What is the experiment objective?	48/50	96%
Action-type Prompts	Valid action identification	45/50	90%
	Invalid action identification	46/50	92%

plex scientific experiments such as "Plant DNA Extraction" and "Phototropism Experiment", which involved higher conceptual complexity. Furthermore, as detailed in Table 2, **question-type prompts** exhibited varied performance depending on the specific query. Objective identification queries performed very well (96% accuracy), indicating the system's robust capability in extracting key experimental information. Phase listing queries showed slightly lower but still strong performance (88% accuracy), suggesting effective procedural understanding. For **action-type prompts**, the system demonstrated balanced performance in both valid action identification (90% accuracy) and invalid action identification (92% accuracy). This indicates reliable capability in procedural validation and error detection, which is crucial for guided experiment execution.

Finally, in order to demonstrate the impact of the ontological guidance, we performed a comparative analysis of responses regarding the "Phalaenopsis Orchid" in a biological experiment context.

Without ontological guidance:

> *The Phalenopsis Orchid is a type of plant belonging to the Orchidaceae family. In the experiment described in the document, it is observed that the cells of the leaf of the Phalenopsis orchid are polygonal and contain chloroplasts and stomata, which are responsible for chlorophyll photosynthesis. For further information about this plant, I recommend consulting an additional source, as the document only provides information related to a specific experiment.*

With ontological guidance:

> *The Phalaenopsis Orchid is a species of orchid that has polygonal cells containing chloroplasts (the site of photosynthesis) and stomata (which regulate gas exchange). These characteristics were observed under the microscope during the experiment described in the document. In particular, the cells of the Phalaenopsis Orchid contain chloroplasts, which are green organelles essential for photosynthesis. This suggests that these cells are part of a photosynthetic tissue exposed to light. Additionally, the cells of the Phalaenopsis Orchid also have stomata, epidermal openings that regulate gas exchanges with the environment. This suggests that these cells are adapted to regulate water and gas balance.*

Without ontological constraints, the system provided basic taxonomic information but acknowledged limitations in domain-specific detail. In contrast, the ontology-guided system delivered complete explanations of cellular structures and their functions, demonstrating enhanced domain-specific knowledge integration and more detailed biological reasoning.

4 Conclusions and Future Work

This paper presented a novel framework that combines LLMs with ontological guidance for the automatic extraction of entities and semantic relationships from

educational and technical manuals. The proposed system addresses the critical challenge of transforming unstructured technical documentation into actionable, structured knowledge that can guide users through complex procedural tasks.

The framework's three-module architecture demonstrates effective integration of retrieval-augmented generation with domain-specific constraints. The experimental evaluation across ten diverse scientific manuals showed strong performance across all three knowledge extraction tasks, with particularly robust results in procedural understanding and factual information retrieval. Furthermore, we showed that ontological guidance significantly enhances the system's ability to provide domain-specific, contextually rich responses.

Future work will focus on expanding the framework to support multi-modal inputs, investigating automated ontology generation techniques, and evaluating the system's effectiveness in real-world educational and industrial settings. Additionally, we plan to explore the integration of user feedback mechanisms to continuously improve the system's performance and adaptability across diverse technical domains.

Acknowledgments. This work was partially supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU and its cascade call project VASARI - Virtual reAlity and Simbiotic ARtificial intelligence in Industrial operation simulation.

References

1. Beauchamp, G., Kennewell, S.: Interactivity in the classroom and its impact on learning. Computers & Education **54**(3) (2010)
2. Hokanson, B., Hooper, S.: Computers as cognitive media: examining the potential of computers in education. Computers in Human Behavior **16**(5) (2000)
3. Kumar, A., Starly, B.: "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. Journal of Intelligent Manufacturing **33**(8) (2022)
4. Lewis, P., Perez, E., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. in Neural Information Processing Systems **33** (2020)
5. Liu, X., Erkoyuncu, J.A., et al.: Knowledge extraction for additive manufacturing process via named entity recognition with LLMs. Robotics and Computer-Integrated Manufacturing **93** (2025)
6. Manesh, M.F., Pellegrini, M.M., et al.: Knowledge management in the fourth industrial revolution: Mapping the literature and scoping future avenues. IEEE Transactions on Engineering Management **68**(1) (2020)
7. Rula, A., Calegari, G.R., et al.: Annotation and extraction of industrial procedural knowledge from textual documents. In: Proceedings of the 12th Knowledge Capture Conf. 2023 (2023)
8. Song, K., Tan, X., et al.: Mpnet: Masked and permuted pre-training for language understanding. Adv. in Neural Information Processing Systems **33** (2020)
9. Wei, F., Neary, R., et al.: Empirical evaluation of embedding models in the context of text classification in document review in construction delay disputes. In: 2024 IEEE Int. Conf. on Big Data (2024)