A Multimodal Approach for Cross-Table Knowledge Graph Construction in Technical Documents

Hao-Ze Wang[1] , Yi-Shin Chen[2]
[1] *Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan*
[2] *Computer Science, National Tsing Hua University, Hsinchu, Taiwan*
*Contact: haoze@gapp.nthu.edu.tw, phone +886-968-300-102

Abstract— **Tables are a fundamental tool in technical documents for presenting structured data. Despite advancements in table understanding, a significant gap remains in handling the specialized, parameter-driven tables commonly found in technical documents. These tables often contain complex technical terms and parameters, posing challenges for large language models (LLMs) and retrieval-augmented generation (RAG) systems, which are typically trained on everyday data. We propose an innovative multimodal approach that combines computer vision (CV) and natural language processing (NLP) to create a dual-layer knowledge graph (KG) for technical table understanding. First, CV techniques process table images to identify headers and rows, creating a cross-table overview by transforming each row into a node. Next, NLP techniques extract and analyze row content, forming detailed semantic connections between individual cell values. This layered KG enables both macro- and micro-level analysis, improving the interpretation of complex technical tables and significantly reducing the time engineers spend retrieving and understanding information in technical documents.**

I. INTRODUCTION

Tables are commonly used in technical documents, such as industry standards and engineering specifications, to present structured data. Unlike general-purpose tables, which often contain rich semantic information, technical tables are filled with parameters, abbreviations, and numeric data that lack semantic cues. This makes it difficult for standard natural language processing (NLP) techniques and large language models (LLMs) to interpret the content effectively. As Sui et al. (2024) [4] demonstrated, LLMs struggle with tasks such as table size detection and merged cell identification, highlighting the need for more specialized methods.

Previous approaches have primarily focused on single-table interpretation and the extraction of structured relationships. For instance, Liu et al. (2023) [1] explored linking table cells to knowledge graphs but did not address the complexity of cross-table relationships, which are crucial in many technical domains. Multimodal methods, like Li et al. (2022) [2], combined computer vision (CV) and NLP for table parsing but were limited to financial datasets rich in semantic information. Technical tables, however, often contain sparse semantic content, requiring more robust approaches.

To address these challenges, we propose a novel multimodal framework that integrates CV and NLP to better understand technical tables. Building on the insights of **Deng et al. (2022)** [3], who focused on relational tables, our approach goes beyond single-table interpretation by constructing a cross-table knowledge graph. This framework captures both the structure and semantic relationships in technical tables, providing both a **macro-level overview** of table relationships across different documents and a **micro-level analysis** of the detailed semantic connections within individual cells. By integrating these two perspectives, our approach enhances the interpretation and retrieval of complex technical data, facilitating more efficient navigation of technical documents.

II. RESEARCH METHODS

A. Phase 1 - Table Parsing & Graph Initialization

A.1 Header and Data Cell Classification

In the first phase, we train a Computer Vision model using the ComFinTab dataset to accurately classify header and data cells in technical tables. The model takes table images as input and distinguishes cells based on visual features such as position and alignment. Each cell ci is classified as either a header or data cell using the function $f_{cv}(c_i)$.

A.2 Cell Localization and Content Extraction

After classifying each cell as either a header or data cell using the Computer Vision model trained on the ComFinTab dataset, we proceed with extracting the content from each cell. Using img2table technology, we convert the visual representation of the table into structured textual data. Each cell C_i is localized by its bounding box coordinates, and the corresponding content t_i is extracted for further processing.

A.3 Constructing the Initial Table Graph

Once the cell contents are extracted, we represent each row as a node in the knowledge graph. The headers of the table are used as attributes, and the corresponding cell values are stored as the values of these attributes. For each node n_i , the attribute set A_j is defined as:

$$A_j = \{(h_1, t_{j1}), (h_2, t_{j2}), \dots, (h_k, t_{jk})\}$$

Where $h_1, h_2, \dots, h_k$ represent the headers, and $t_{j1}, t_{j2}, \dots, t_{jk}$ are the data values from each row. This creates a graph structure

where each row is connected to its attributes, representing the core structure of the table.

To establish relationships between rows (nodes), we compute cosine similarity between their content. If the similarity score $s(n_j, n_k)$ exceeds a predefined threshold, we create an edge between the nodes.

At this stage, we have constructed a **macro-level** knowledge graph that captures the structural relationships between rows in the table, providing a high-level representation of the table's organization. This forms the basis for more detailed semantic analysis in the next phase.

B. Phase 2 - Semantic Analysis and Refinement

B.1 Entity Extraction and Relationship Identification

In this phase, the nodes in the knowledge graph represent key **entities** extracted from technical documents, such as parameters or specialized terms, rather than entire rows. Using **NER** and **LLM** (e.g., Llama 3.2), we identify these entities and the relationships between them from the cell content.

B.2 Semantic Similarity Between Nodes

Once the entities and relationships are extracted, we compute the **semantic similarity** between these entities. Each entity e_j and e_k is embedded into a vector space using the LLM to generate embeddings E_j and E_k. The semantic similarity score $s_{sem}(e_j, e_k)$ between two entities.

Entities with a high semantic similarity score are connected in the graph, reflecting meaningful relationships between them. This enables the graph to capture both explicit and implicit relationships between specialized terms and parameters at a **micro-level**, going beyond surface text similarity to reveal deeper contextual connections through the LLM.

III. Experimental Evaluation

We evaluate our method, **Cross Table KG**, for cross-table knowledge extraction, focusing on the accuracy of retrieving row-based information from tables in ISO documents. The models are tested on queries that require correctly identifying rows across multiple tables to answer specific technical questions.

1. Dataset

The evaluation is conducted on two related ISO documents containing over 300 tables. The test set includes 2,000 cross-table questions designed to assess the models' ability to extract and link information across tables. These questions require the correct identification of rows and the retrieval of relevant information from different tables within the documents.

2. Comparison Models

- **GPT-4.0**: A general-purpose LLM tested on the two ISO document PDFs.
- **LLAMA 3.2** (11B, 90B): Different versions of LLAMA are evaluated to understand the impact of model size on cross-table queries.
- **CLEAR**: Our base model, which hierarchically structures PDFs but lacks specific table understanding and linking capabilities.
- **Our Method (Cross Table KG)**: Builds on CLEAR with the addition of cross-table knowledge graph (KG) functionality, designed to enhance table linking and row-based information extraction.

3. Evaluation Metrics

- **Cross-table Accuracy**: Measures the correctness of information retrieval that spans multiple tables.
- **Row-based Precision**: Assesses the accuracy of identifying the correct row when extracting information.
- **Row-based Recall**: Measures the model's ability to retrieve all relevant rows in cross-table queries.

TABLE I
PERFORMANCE COMPARISON OF MODELS ON CROSS-TABLE QUERIES

Model	Cross-table Accuracy (%)	Row-based Precision (%)	Row-based Recall (%)
GPT-4.0	61.21	71.21	69.23
LLAMA 3.2 11B	52.71	62.51	60.61
LLAMA 3.2 90B	58.21	69.87	67.97
CLEAR	43.52	35.76	33.21
Cross Table KG	**86.78**	**88.23**	**87.41**

4. Results

Our method, **Cross Table KG**, outperformed the other models in both accuracy and row identification. **GPT-4.0** and **LLAMA models** struggled with complex table linking, while **CLEAR** was unable to accurately extract row-based information without table-specific understanding. The addition of cross-table knowledge graph construction in our method resulted in significantly higher accuracy and precision across all tested queries.

IV. Conclusions

Our proposed multimodal framework significantly improves the interpretation and retrieval of technical table data by constructing both macro- and micro-level knowledge graphs. By leveraging CV for structural extraction and LLM for semantic analysis, we capture both explicit and implicit relationships between specialized terms, enabling more efficient navigation of complex technical documents. This approach addresses the limitations of existing methods and provides a more robust solution for understanding parameter-driven technical tables.

References

[1] Liu, Jixiong, et al. "From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods." Journal of Web Semantics 76 (2023): 100761.

[2] Li, Zaisheng, et al. "End-to-End Compound Table Understanding with Multi-Modal Modeling." Proceedings of the 30th ACM International Conference on Multimedia. 2022.

[3] Deng, Xiang, et al. "Turl: Table understanding through representation learning." ACM SIGMOD Record 51.1 (2022): 33-40.

[4] Sui, Yuan, et al. "Table meets llm: Can large language models understand structured table data? a benchmark and empirical study." Proceedings of the 17th ACM International Conference on Web Search and Data Mining. 2024.