Scientific Table Data Extraction with Uncertainty Quantification

Kehinde Ajayi
kajay001@odu.edu
Old Dominion University
Norfolk, Virginia, USA

Abstract

Complex scientific tables present unique challenges for information extraction due to their multi-level headers, merged cells, and domain-specific notations. Existing Table Structure Recognition (TSR) frameworks, often fall short when applied to these complex structures. How to perform UQ effectively and efficiently for table data extraction is a research question. To address these gaps, we propose an integrated pipeline that leverages artificial intelligence (AI) methods for mining complex scientific tables. Our approach combines TSR, Optical Character Recognition (OCR), and Large Language Models (LLMs) with uncertainty quantification techniques. We introduce the GenTSR benchmark for evaluating TSR methods across scientific domains and a modified Test-Time Augmentation (TTA-m) approach for uncertainty quantification. Additionally, we propose a novel benchmark for LLM-based table question-answering tasks using complex scientific tables. This comprehensive framework aims to enhance the accuracy and reliability of information extraction from scientific tables, facilitating more effective data analysis and interpretation in various research domains.

CCS Concepts

• **Information systems → Information retrieval**.

Keywords

Table Structure Recognition, Table Data Extraction, Uncertainty Quantification, Large Language Models

ACM Reference Format:
Kehinde Ajayi. 2024. Scientific Table Data Extraction with Uncertainty Quantification. In *The 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL '24), December 16–20, 2024, Hong Kong, China.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3677389.3702616

1 Introduction

Scientific tables are essential for representing experimental data, results, and key findings in academic and technical documents. Extracting structured data from these tables has been a central problem in document analysis and information retrieval for decades. Table Structure Recognition (TSR), which involves identifying the rows, columns, and cells of tables from images or digital documents, is a critical task in this domain. However, despite significant advancements, TSR methods still face challenges in handling complex table structures found in scientific documents. Moreover, existing methods rarely offer measurements of uncertainty in their predictions, limiting the data verification in downstream tasks such as data analysis, modeling, and decision-making [15].

Early approaches to TSR used rule-based or heuristic methods to extract table structures, which were often tailored to specific types of documents [3]. These methods lacked generalizability and failed on complex tables as appeared in scholarly papers. In recent years, deep learning-based approaches have become the state of the art. Models such as CascadeTabNet [12] and SPLERGE [17] use convolutional neural networks (CNNs) and transformers to detect tables and extract their structures from document images. While these models achieve high accuracy in detecting rows, columns, and cells, they do not quantify uncertainties, which is critical for data validation. This limitation is particularly significant in scientific documents, where precision is paramount.

Uncertainty quantification (UQ) methods have gained traction in various machine learning domains, including computer vision and natural language processing (NLP) [9]. In the context of TSR, UQ can provide confidence scores for extracted table structures, allowing users to assess the reliability of the extracted information. To address this problem, We introduced UQ in TSR, using Test-Time Augmentation (TTA) to estimate uncertainties in TSR outputs [2]. However, the existing study was limited to specific TSR models and datasets, leaving room for further exploration of more general UQ methods, such as Conformal Prediction [14].

OCR methods like PaddleOCR [5] and General of Theory (GOT) [18], have advanced the extraction of text from table images. PaddleOCR focuses on detecting bounding boxes and extracting text data, while GOT can extract table content in LaTeX format, making it particularly suitable for scientific tables. However, integrating OCR outputs with TSR models remains a challenge, particularly in handling complex table layouts in scientific documents [13]. Additionally, LLMs, such as GPT-4, have shown promise in interpreting table data and answering questions about it, but their application in table QA tasks is still in its infancy [10].

My PhD thesis aims to bridge these gaps by integrating advanced TSR, OCR, and LLM methods with UQ techniques to provide a comprehensive solution for extracting and understanding table data from scientific documents. We propose the use of Conformal Prediction to quantify uncertainties in both table structure and content extraction tasks, as well as a new benchmark for evaluating LLM-based table question answering (QA) tasks using complex scientific tables. In addition, we propose a framework that leverages multi-agent LLMs to improve table data extraction.

JCDL '24, December 16–20, 2024, Hong Kong, China

ACM ISBN 979-8-4007-1093-3/24/12
https://doi.org/10.1145/3677389.3702616

2 Research Questions

The research questions (RQs) for this study are as follows:

- **RQ 1**: Reproducibility and replicability are critical for ensuring reliable TSR models. What is the status of reproducibility and replicability of existing TSR models?
- **RQ 2:** Quantifying the uncertainties in table structures can increase the reliability of TSR methods. How can we develop a pipeline to accurately quantify the uncertainties of TSR models?
- **RQ 3:** Quantifying the uncertainties in both table structures and table cell contents can improve the efficiency of data verification so human validators can focus on only errors in extracted data. How can we develop a pipeline that Integrates TSR, OCR, and UQ to improve image-based table data extraction accuracy and confidence?
- **RQ 4:** What is the performance of state-of-the-art commercial and open-weight LLMs on complex table question answering?

3 Methodology

3.1 Preliminary Work

Data Collection To answer RQ 3, we annotated 200 table images from PDFs in 5 scientific domains (Material Science, Biology, Computer Science, Scientific Reports, and ICDAR 2013) using the VGG Image Annotator (VIA) [6]. VIA is an open-source software for annotating images, videos, and audio. We drew rectangular bounding boxes around text content in a table cell and provided properties including "start-row", "start-col", "end-row", and "end-col" as labels. We used the Amazon Textract tool to obtain the cell contents and included a "text" label to the properties above. To answer RQ 4, we will build a new dataset comprising complex scientific tables and LLMs-generated questions, extending traditional QA benchmarks that rely on simpler Wikipedia tables [9]. This dataset will be used to assess the reasoning and interpretative capabilities of models such as GPT-3.5 [10].

Reproducibility and Replicability of TSR Methods To investigate the reproducibility and replicability of different TSR methods across different datasets, we introduced a benchmark, GenTSR, which consists of 386 table images obtained from research papers in six scientific domains, including three STEM and three non-STEM domains. We manually annotated GenTSR using the VIA [6] following the same schema as the ICDAR 2019 dataset. Our Reproducibility tests evaluate models on original datasets, while replicability tests use alternate or GenTSR datasets, with F-scores computed at five IoU thresholds from 0.5 to 0.9. [1]. Our reproducibility experiment shows that 4 [7, 8, 19, 20] out of 6 executable TSR methods were labeled reproducible, 1 paper [12] was labeled partially-reproducible, and 1 paper was labeled not-reproducible [17]. None of the 4 methods that allow inference on custom data [7, 8, 12, 17] was replicable with respect to the GenTSR dataset, under a threshold of 10% absolute F-score.

Uncertainty Quantification in TSR Methods To ensure that the outputs of TSR methods are reliable, we proposed a novel pipeline for UQ in TSR using a modified Test-Time Augmentation (TTA) approach called TTA-m [2]. Our UQ pipeline consists of 4 components: data augmentation, TSR model fine-tuning, TTA, and confidence estimation via ensembles. We evaluated our pipeline using the ICDAR 2019 modern dataset. To assess the effectiveness of our UQ method, we introduced two heuristics: masking and cell complexity quantification. Our results showed that the TTA-m model outperformed baseline methods in terms of F1 scores for cell detection. Additionally, we found that our method accurately captured increased uncertainty when table image pixel intensity was decreased and when cell complexity (measured by adjacency degrees) increased.

3.2 Proposed Work

UQ on Table Data Extraction The existing SOTA models (e.g., TableNet [11]) for table understanding implement both table detection (TD) and TSR on table images. However, none of these methods incorporates table content extraction via OCR nor quantify the uncertainties in the extracted data. To address this problem, we propose a pipeline that performs UQ on table data extraction obtained via:

(1) **The integration of TSR with OCR**. Specifically, we will use the Table Transformer model [16] to obtain the row-column information for the table cells and text contents using PaddleOCR [5]. These two pieces of information will be combined to provide row and column identification.
(2) **The integration of OCR and LLM**. We will use General of Theory (GOT) [18], a transformer-based OCR model that returns table cell text in LaTeX format. This LaTeX output will be passed to an LLM, fine-tuned on LaTeX tabular data, to provide both table cell locations and extracted text.
(3) **Exploration of UQ Methods** To quantify the uncertainties in the above extraction results, we will explore and compare several UQ methods such as the conformal prediction method.

Multi-agent LLM for improved table data extraction We propose a multi-agent LLMs that will leverage both image and text modalities to improve the accuracy of table data extraction results.

LLM-based Table QA Benchmark: Current benchmarks for Table QA tasks consist of tables from Wikipedia [4] and non-scientific domains, which do not reflect the complexities found in real-world tables. In addition, these datasets consist of questions and answers manually created by human experts, which can be very expensive. To overcome these problems, we propose a complex scientific table QA benchmark consisting of tables from the ICDAR, Material Science, Biology, Computer Science, and Scientific Reports domains. We will use LLMs such as GPT-3.5, Llama 2 and 3, and Mistral-7B to generate questions, supply answers, and provide explanations, offering a comprehensive evaluation of their reasoning capabilities. We will evaluate each LLM on questions created by other LLMs.

4 Conclusion

We will develop, a library that is capable of integrating UQ into scientific table extraction and understanding tasks. By implementing UQ in combination with TSR, OCR, and LLM methods including multi-agent LLMs, we aim to provide more reliable extraction results, with uncertainty scores indicating confidence in the output.

References

[1] Kehinde Ajayi, Muntabir Hasan Choudhury, Sarah M Rajtmajer, and Jian Wu. 2023. A Study on Reproducibility and Replicability of Table Structure Recognition Methods. In *International Conference on Document Analysis and Recognition*. Springer, 3–19.

[2] Kehinde Ajayi, Leizhen Zhang, Yi He, and Jian Wu. 2024. Uncertainty Quantification in Table Structure Recognition. In *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 1–6.

[3] Florence Folake Babatunde, Bolanle Adefowoke Ojokoh, Samuel Adebayo Oluwadare, et al. 2015. Automatic table recognition and extraction from heterogeneous documents. *Journal of Computer and Communications* 3, 12 (2015), 100.

[4] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2013. Methods for exploring and mining tables on wikipedia. In *Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics*. 18–26.

[5] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941* (2020).

[6] Abhishek Dutta and Andrew Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*. 2276–2279.

[7] Pascal Fischer, Alen Smajic, Giuseppe Abrami, and Alexander Mehler. 2021. Multi-Type-TD-TSR–Extracting Tables from Document Images Using a Multi-stage Pipeline for Table Detection and Table Structure Recognition: From OCR to Structured Table Representations. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 95–108.

[8] Eunji Lee, Jaewoo Park, Hyung Il Koo, and Nam Ik Cho. 2022. Deep-learning and graph-based approach to table structure recognition. *Multimedia Tools and Applications* 81, 4 (2022), 5827–5848.

[9] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2ORC: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782* (2019).

[10] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2, 5 (2023).

[11] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2019. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 128–133.

[12] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 2020. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 572–573.

[13] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. 1162–1167. https://doi.org/10.1109/ICDAR.2017.192

[14] Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research* 9, 3 (2008).

[15] Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. 2019. DeepTabStR: deep learning based table structure recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1403–1409.

[16] Brandon Smock and Rohith Pesala. 2021. *Table Transformer*. https://github.com/microsoft/table-transformer

[17] Chris Tensmeyer, Vlad I. Morariu, Brian Price, Scott Cohen, and Tony Martinez. 2019. Deep Splitting and Merging for Table Structure Decomposition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 114–121. https://doi.org/10.1109/ICDAR.2019.00027

[18] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model. *arXiv preprint arXiv:2409.01704* (2024).

[19] Wenyuan Xue, Qingyong Li, and Dacheng Tao. 2019. ReS2TIM: Reconstruct Syntactic Structures from Table Images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 749–755. https://doi.org/10.1109/ICDAR.2019.00125

[20] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. 2021. TGRNet: A Table Graph Reconstruction Network for Table Structure Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1295–1304.