A Study on Data Extraction Techniques for Semi-Structured Scientific Documents

Abstract

This paper presents an analysis of text extraction methods applied to semi-structured scientific documents, focusing on PDF and DOCX formats. We evaluate a range of heuristics including heading detection, caption extraction, and reference linking. Our results show that hybrid rule-based approaches outperform naive text extraction in cases where document structure is partially preserved. The study aims to assist researchers developing automated metadata generation pipelines.

1. Introduction

Scientific documents are commonly distributed in PDF and DOCX formats, each presenting unique challenges for automated extraction. PDFs often lose logical reading order, while DOCX files retain semantic structure such as headings and styled paragraphs.

This paper evaluates methods to reliably extract headings, captions, references, and other metadata components from both formats.

2. Related Work

Previous studies have explored layout-aware extraction, natural language processing pipelines, and machine learning models for document segmentation. However, limited work focuses on lightweight, rule-based approaches suitable for small-scale research workflows.

Recent advancements in agentic AI systems offer new opportunities for semi-automated document processing pipelines.

3. Methodology

Our approach evaluates three categories of extraction techniques: rule-based heuristics, regex-driven caption detection, and block-based text grouping for PDFs.

Figure 1 illustrates the overall extraction pipeline used in our experiments. Table 1 summarizes the documents included in the dataset.

4. Results

Figure 2 highlights the accuracy comparison between PDF and DOCX extraction outputs. We observed that DOCX-based extraction consistently outperformed PDF-based extraction due to richer structural metadata.

Table 2 presents performance metrics including precision and recall for heading and caption detection tasks.

5. Discussion

While rule-based methods are interpretable and lightweight, they fail in cases where documents follow unconventional layouts. This suggests the need for hybrid techniques that combine rule-based logic with ML-driven classification.

Future work includes integrating transformer-based models to detect semantic sections and improve caption-reference linking accuracy.

6. Conclusion

This study demonstrates the effectiveness of structure-aware extraction heuristics for scientific documents. Our findings provide a foundation for building scalable, automated metadata generation pipelines.

Figures

Figure 1: Overview of the extraction pipeline architecture.

Figure 2: Accuracy comparison between PDF and DOCX extraction techniques.

Tables

Table 1: Dataset summary including document counts and formats.

Table 2: Performance metrics for heading and caption detection tasks.