Research Report

Shahzada Muhammad Ali

Structure-Aware Parsing and Knowledge Graph Construction from Complex Legal Documents

Abstract
Legal texts such as penal codes exhibit deep hierarchical structure, dense cross-referencing, and layout-dependent semantics that are poorly handled by standard document parsing and retrieval pipelines. Existing approaches that rely on HTML sources or flat text extraction often lose section boundaries, footnotes, and legal context, limiting their applicability for downstream reasoning tasks such as retrieval-augmented generation (RAG) and knowledge graph construction. In this work, we present a structure-aware parsing pipeline that reconstructs chapters, sections, and footnotes directly from authoritative PDF sources using layout metadata, symbolic parsing, and heuristic alignment. We demonstrate how font-level signals and incremental correction strategies can recover reliable legal structure in the presence of noisy or incomplete sources, and we produce a graph-ready representation suitable for legal reasoning and GraphRAG systems.

1. Introduction

Legal documents pose unique challenges for natural language processing due to their strict structural requirements and tolerance for neither ambiguity nor hallucination. Penal codes, in particular, are organized hierarchically into chapters, sections, clauses, and footnotes, where meaning often depends on exact numbering, cross-references, and exceptions embedded in footnotes.

Most existing pipelines for legal document processing rely on either:

1. HTML versions of statutes, which are frequently incomplete, inconsistent, or editorialized, or
2. Plain-text extraction from PDFs, which discards layout information critical to recovering structure.

These limitations become especially problematic for downstream applications such as legal RAG, statutory reasoning, and knowledge graph construction, where structural fidelity is essential.

This work addresses the following research question:
"How can we reliably recover the semantic and hierarchical structure of complex legal documents directly from authoritative PDFs, in a manner suitable for graph-based reasoning and retrieval?"

2. Problem Definition

Given a legally authoritative PDF of the Pakistan Penal Code, the objective is to construct a machine-usable representation that preserves:

- Chapter and section boundaries
- Section titles and numbering (including alphanumeric sections)
- Section body text
- Footnotes and their linkage to relevant sections
- Ordering and hierarchy required for graph construction

The problem is constrained by:

- Inconsistent formatting across chapters
- Section merges, omissions, and numbering anomalies
- Footnotes that are visually encoded rather than semantically marked
- The unreliability of HTML sources as a single ground truth

3. Related Challenges

Unlike general document parsing, legal parsing must handle:

- **Layout-driven semantics**: font size, capitalization, and placement encode meaning.
- **Non-local dependencies**: footnotes modify interpretation across sections.
- **High precision requirements**: incorrect section boundaries invalidate reasoning.

This motivates a structure-first, layout-aware approach rather than end-to-end statistical parsing.

4. Methodology

We implement a multi-stage parsing pipeline that combines PDF layout analysis, symbolic parsing, heuristic alignment, and incremental correction.

4.1 PDF Text and Layout Extraction

We use PyMuPDF to extract text along with font metadata (font name, size, capitalization, and style). Unlike OCR-only pipelines, this preserves layout signals that are critical for identifying chapters, section titles, and footnotes. Non-semantic artifacts such as page numbering ("Page X of Y") are removed during preprocessing.

4.2 Chapter Detection

Chapters are detected using a combination of:

- Roman numeral patterns (e.g., *CHAPTER XIV*)
- Capitalization heuristics for titles
- Positional constraints relative to chapter markers

Where chapter titles are missing or malformed, corrective logic is applied based on domain knowledge of the legal text. This reflects a realistic research setting in which authoritative sources are imperfect.

4.3 Section Identification and Title Reconstruction

Sections are identified using regular expressions that capture:

- Numeric sections (e.g., *228*)
- Alphanumeric sections (e.g., *337Y*, *365A*)

Section titles are reconstructed by aggregating contiguous lines until a structural boundary (next section or chapter) is encountered. To address merged or malformed sections, targeted correction strategies are applied. For example, merged section titles are split based on known legal patterns, and missing sections are inserted explicitly when authoritative titles are known.

4.4 Section Text Alignment

To recover section bodies, the entire document is normalized into a cleaned text stream. For each section:

1. The section title is normalized and searched in the cleaned text.
2. Text between consecutive section titles is extracted.
3. Fuzzy string matching and numeric proximity checks are used to validate alignment.

This process resembles weakly supervised alignment, relying on symbolic and layout cues rather than annotated data.

4.5 Footnote Detection and Linking

Footnotes are handled using two complementary strategies:

Inline footnotes are detected by identifying small-font numeric spans followed by explanatory text.

Footnote-bar entries are extracted by scanning page footers for low-font-size text sequences initiated by numeric markers.

Footnotes are indexed by page-local identifiers to preserve provenance. Duplicate or fragmented footnotes are merged conservatively. Long or non-essential footnotes are intentionally excluded from the final representation unless semantically necessary, reflecting precision-driven design choices common in legal NLP.

5. Hybrid Source Validation

HTML versions of the penal code are explored as a secondary source. However, due to structural omissions and inconsistent markup, HTML is used only for validation and fallback. The PDF remains the primary authoritative source. This hybrid strategy allows cross-checking without sacrificing structural integrity.

6. Knowledge Graph Readiness

The final output is a hierarchical JSON structure:

Chapter → Section → {title, text, footnotes}

This representation directly supports:

- Node-level section indexing
- Relation-only graph construction
- GraphRAG pipelines
- Cypher-based querying with LLM agents

Preliminary agent tooling demonstrates how parsed sections can be queried safely using parameterized graph queries.

7. Discussion

This work demonstrates that high-fidelity legal parsing is achievable without large labeled datasets by exploiting layout signals and domain structure. The pipeline emphasizes interpretability, correctness, and incremental validation—properties that are critical in legal AI systems.

Rather than treating parsing errors as noise, the system explicitly models and corrects them, reflecting a research-oriented approach to real-world data.

8. Limitations and Future Work

Current limitations include:

- Manual correction logic for rare section anomalies
- Heuristic thresholds for font-based detection
- Limited semantic interpretation of footnotes

Future work includes:

- Learning layout-to-structure mappings using weak supervision
- Integrating semantic role extraction for legal relations

- Coupling the parser with GraphRAG-based reasoning models
- Extending the pipeline to case law and multi-jurisdictional statutes

9. GenAI Usage Disclosure

We employed ChatGPT to assist in rephrasing the report for improved clarity. All core content, including research design, data analysis, and result interpretation, was conducted without the aid of generative AI tools.