Reconstruction of Text Hierarchy in Legal Documents

Project Plan

Meeran Mydeen Syed Ibrahim- 223201700
Seshant Babu Kodamunja -223201732
Ashwin Rajendran Appuraj-224100211
Karamchand Subash Kamaraj - 224100274

Research Questions

1.How can hierarchical document structure analysis (e.g., headings, paragraphs, tables) be accurately extracted from scanned vocational education documents with OCR errors?

2.What methods are most effective for reconstructing logical document hierarchies in noisy, semi-structured documents?

3.How can a custom TEI XML schema be designed to represent the extracted structure while preserving semantic relationships?

Aims

- **Literature Review**: Survey existing methods for document structure recognition (e.g., Detect-Order-Construct, HELD, rule-based approaches).
- **Dataset Creation**: Annotate 280 scanned BIBB documents with structural labels (headings, paragraphs, lists) in TEI XML format.
- **Model Development**: Train and fine-tune a multimodal AI model for structure recognition.
- **Evaluation**: Assess model performance on layout analysis, reading order, and hierarchical reconstruction.

Work Packages

WP1: Literature Review & Methodology Selection

- Review existing approaches (e.g., Detect-Order-Construct, HELD, DocParser).
- Identify best practices for handling OCR errors and layout inconsistencies.
- Select a hybrid approach combining vision (CNN) and text (Transformer) models.

WP2: Dataset Annotation & TEI XML Schema Design

- Define annotation guidelines (e.g., headings, paragraphs, lists, tables).
- Manually annotate 280 BIBB documents in TEI XML format.
- Validate annotations for consistency and correctness.

WP3: Model Training & Optimization

- Preprocess data (OCR correction, layout normalization)
- Fine-tune a multimodal model (e.g., LayoutLMv3, Detect-Order-Construct framework)
- Optimize for hierarchical structure prediction (e.g., reading order, TOC extraction)

WP4: Evaluation & Error Analysis

- Benchmark against rule-based and existing deep learning baselines
- Evaluate using metrics F1-score, reading order accuracy etc

WP5: Documentation & Reporting

- Prepare a technical report on methodology, results, and limitations

Tasks

Task 1: Literature Review (Weeks 1–2)

- Summarize key papers on document structure analysis

Task 2: TEI XML Schema Design (Weeks 3–4)

- Define XML tags for document elements (e.g., <div>, <head>, <p>, <list>)

Task 3: Manual Annotation and Data Preprocessing (Weeks 5-7)

- Annotate 280 documents using tools like BRAT or Prodigy
- Clean OCR errors

Task 4: Model Training (Weeks 8-11)

- Implement Detect-Order-Construct or LayoutLMv3
- Train on 80% of data, validate on 10%, test on 10%

Task 5: Evaluation (Weeks 12-13)

- Compare model predictions against ground truth
- Compute metrics

Task 6: Final Report (Weeks 14-15)

- Summarize findings, limitations, and future work.

Milestones

Milestone	Deadline	Deliverable
TEI XML schema finalized	Week 4	Schema document
Dataset fully annotated	Week 7	Annotated XML files
Baseline model trained	Week 11	Model checkpoint
Final model evaluation completed	Week 13	Performance metrics
Project report submitted	Week 15	Final report

Expected Challenges

- OCR errors
- Inconsistent document layouts
- Complex hierarchies
- TEI XML schema flexibility

Conclusion

- This project plan outlines a structured approach to extracting hierarchical document structures from vocational education texts. By leveraging multimodal AI models and a carefully annotated TEI XML dataset, we aim to improve automated text structure recognition despite OCR and layout challenges
- This initial plan is subject to change in future depending upon the feasibility of currently mentioned techniques
- The changes will be updated from time to time as we proceed ahead
- The final deliverables will include an annotated dataset, trained models, and a comprehensive evaluation report