KAP: MLLM-assisted OCR Text Enhancement for Hybrid Retrieval in Chinese Non-Narrative Documents

Hsin-Ling Hsu, Ping-Sheng Lin, Jing-Di Lin, Jengnan Tzeng*
National Chengchi University

Introduction

Hybrid retrieval struggles with Traditional Chinese non-narrative documents due to OCR noise, structural distortions, and poor synonym coverage. We propose Knowledge-Aware Preprocessing (KAP), a two-stage framework using Multimodal LLMs to correct OCR errors, restore layout, and optimize text for both sparse and dense retrieval.
Our code is available at github.com/JustinHsu1019/KAP.

Method

Our Knowledge-Aware Preprocessing (KAP) framework employs a two-stage approach to enhance text from non-narrative documents. Stage one utilizes Tesseract OCR for initial text extraction from PDFs in Traditional Chinese. The second stage, MLLM Post-OCR Processing, leverages the Claude 3.7 Sonnet model, integrating both OCR-extracted text and original PDF images. Guided by prompt engineering, this stage encompasses: (1) Error Correction, rectifying OCR inaccuracies;(2) Layout-Aware Format Reconstruction, where MLLM's vision capabilities restore tabular structures and formatting; and (3) Retrieval-Aware Rewriting, optimizing text for BM25 via synonym expansion and for dense retrieval by converting tables to natural language descriptions. Finally, a page-level segmentation followed by recursive chunking (8,000-token chunks, 500-token overlap) is applied to the processed text.

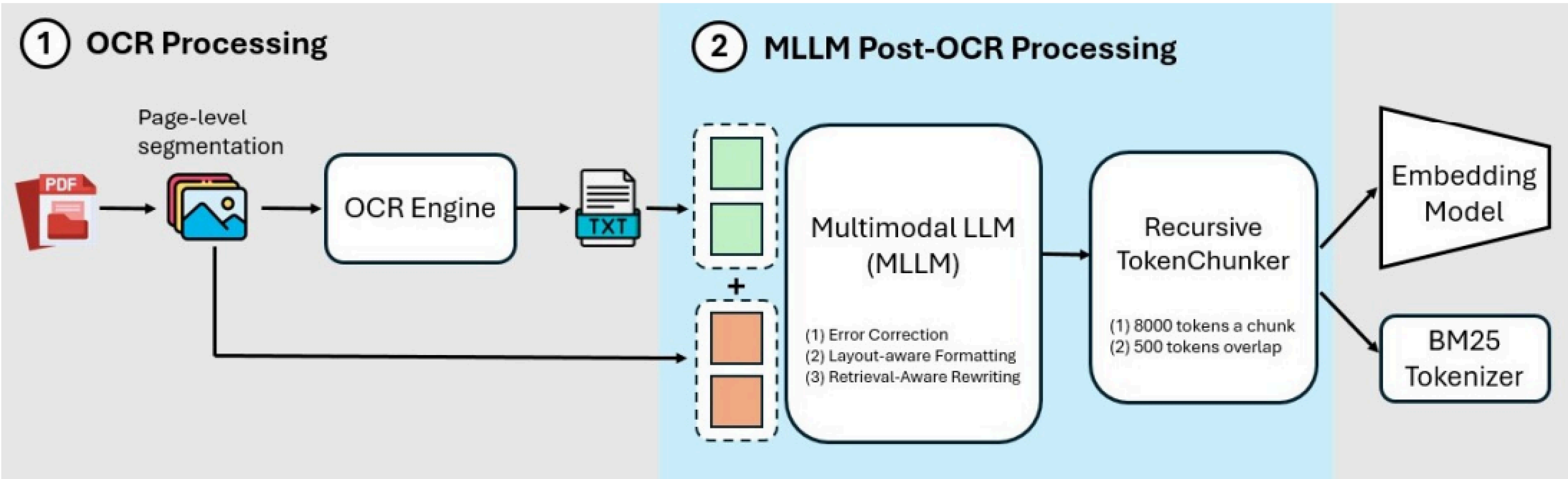

Figure 1: Overall architecture of the proposed KAP framework.

Result

On the E.SUN Bank dataset, KAP significantly outperformed the baseline across all retrieval settings. For Sparse Retrieval (BM25), it improved MRR from 53.16% to 63.64% and Precision@1 from 41.51% to 51.16%. In Dense Retrieval, MRR rose from 48.41% to 65.16% and Precision@1 from 32.10% to 53.65%. Hybrid Retrieval performed best overall, with MRR increasing from 53.23% to 69.46% and Precision@1 from 38.98% to 59.73%.

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	53.16±0.83	41.51±1.67
KAP w/o Vision	54.84±1.24	43.66±1.45
KAP w/o OCR Text	62.32±0.43	49.39±0.73
KAP w/o Rewrite	59.60±1.03	45.45±1.56
KAP (Ours)	**63.64±0.09**	**51.16±0.21**

Figure 2: Performance of Sparse Retrieval

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	48.41±0.60	32.10±0.74
KAP w/o Vision	56.62±0.39	42.98±0.65
KAP w/o OCR Text	54.00±0.63	44.41±0.47
KAP w/o Rewrite	58.46±1.32	46.11±1.65
KAP (Ours)	**65.16±1.51**	**53.65±2.24**

Figure 3: Performance of Dense Retrieval

Methods	MRR (%)	Precision@1 (%)
Tesseract OCR (Baseline)	53.23±0.57	38.98±0.88
KAP w/o Vision	58.52±0.51	47.33±0.81
KAP w/o OCR Text	65.06±0.15	56.39±0.11
KAP w/o Rewrite	66.02±1.71	55.48±2.13
KAP (Ours)	**69.46±0.61**	**59.73±1.10**

Figure 4: Performance of Hybrid Retrieval

Conclusion

KAP significantly enhances text quality and hybrid retrieval accuracy for Traditional Chinese non-narrative documents. This work validates that: (1) MLLM-based post-OCR processing effectively corrects OCR errors and restores critical table structures; (2) our retrieval-aware rewriting module optimizes text representations for both sparse and dense retrieval methods; and (3) these input-level enhancements boost performance without necessitating modifications to existing retrieval architectures. Ablation studies further underscore the vital contribution of each KAP component to the overall observed improvements in retrieval effectiveness.

Acknowledgement

This study was supported by E.SUN Bank, which provided the dataset from the "AI CUP 2024 E.SUN Artificial Intelligence Open Competition".

Hsin-Ling: justin.hsu.1019@gmail.com
Ping-Sheng: guraaaashark@gmail.com
Jing-Di: 111301029@g.nccu.edu.tw
Jengnan: glophy@g.nccu.edu.tw