天主教輔仁大學資訊工程研究所碩士論文

Enhancing Biomedical Document Analysis with Layout-Aware Multimodal model

指導教授：徐嘉連 教授

研究生：杜弘仁 撰

中華民國一百一十二年一月

Abstract

The burgeoning volume of unstructured medical diagnostic documents presents a significant challenge for healthcare systems worldwide. This thesis proposes an innovative end-to-end system leveraging Optical Character Recognition (OCR) and Natural Language Processing (NLP) technologies to automate the extraction of pertinent information from medical diagnoses. By integrating advanced OCR techniques with sophisticated NLP algorithms, the system aims to transform unstructured text into structured data, thereby enhancing the accuracy and efficiency of medical data processing. The evaluation of the system's performance on diverse medical documents underscores its potential to significantly improve clinical decision-making and administrative workflows. This research contributes to the fields of healthcare informatics and artificial intelligence by demonstrating the feasibility and benefits of automating medical information extraction.
keywords： Bert , LayoutXLM , PP-OCRv3.

Contents

Abstract **I**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Background . . . 1
1.3 Objectives . . . 2

2 Related Work **4**

3 Methodology **5**
3.1 Definition . . . 5
3.2 Data Collection . . . 6
3.2.1 Doctor-Patient Information . . . 6
3.2.2 Hospital Information . . . 6
3.2.3 Diagnose-Related Data . . . 6
3.2.4 Doctor Comments Generating . . . 6
3.2.5 Image Rendering . . . 7
3.3 OCR model . . . 8
3.3.1 Text Detection . . . 8
3.3.2 Text Recognition . . . 8
3.4 Key Information Extraction . . . 9
3.4.1 Threshold-Based YX Sorting Algorithm . . . 9
3.5 Mutual Learning . . . 10
3.6 Relational Extraction Model . . . 12

4 Experiment **13**
4.1 Experiment Setup . . . 13
4.1.1 Dataset . . . 13
4.1.2 OCR Text Recognition . . . 13
4.1.3 Relation Extraction . . . 13
4.2 Results . . . 14
4.2.1 OCR Model Performance . . . 14
4.2.1.1 Text Detection and Recognition Performance . . . 14
4.2.2 Key Information Extraction Model . . . 14
4.2.3 Relational Extraction Model . . . 14
4.2.4 End to End System . . . 15

5 Limitations and Future Work **16**

6 Conclusion **16**

1 Introduction

1.1 Motivation

This research addresses the critical challenges in hospital diagnostics by proposing an innovative approach to digitize and manage diagnostic data more efficiently. Leveraging advancements in OCR (Optical Character Recognition) technology and artificial intelligence, it aims to automate the extraction and processing of diagnostic information from paper-based records, thereby enhancing accuracy, reducing manual labor, and improving access to patient data. This solution promises not only to streamline hospital operations but also to significantly improve patient care by allowing healthcare professionals to focus more on treatment decisions and less on administrative tasks.

1.2 Background

This thesis presents a comprehensive, OCR-based system designed to automate the extraction of crucial information from documents, thus enhancing organizational efficiency by minimizing manual interventions and reducing error frequencies. At the heart of this system is the transformation of document images into structured, accessible data. This transformative process commences with the application of Optical Character Recognition (OCR) technology to capture both the text and its spatial positioning within the documents.

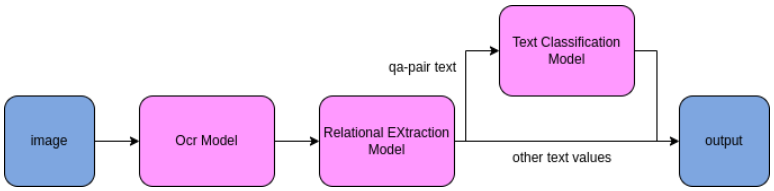


Fig. 1: [?])

Subsequent to this initial text and location capture, the extracted data is funneled into a relational extraction model. The primary objective of this model is to discern and categorize text elements into pairs that share a contextual relationship, thereby illuminating the complex interconnections among various data points within the document. These identified relational pairs are then subjected to analysis by a BERT (Bidirectional Encoder Representations from Transformers) model, specifically tailored for Named Entity Recognition (NER) tasks. This analytical phase is pivotal for the precise identification and extraction of essential information fields.

In parallel, text elements that do not adhere to these identified relational pairs are directly forwarded to the output stage, ensuring a holistic approach to data capture and utilization. This encompasses both the information situated within identified relationships and that which exists outside these

parameters. Through the deployment of advanced computational techniques, this system significantly streamlines the document processing workflow, thereby propelling productivity forward.

1.3 Objectives

The primary aim of this research is to enhance Named Entity Recognition (NER) accuracy through a multimodal approach, leveraging both semantic content and spatial information from documents. This objective is pursued through a series of strategic methodologies:

- Utilization of state-of-the-art Optical Character Recognition (OCR) technology to capture both the textual content and the associated bounding boxes of entities within documents. This dual extraction method is crucial for enriching the textual data with spatial context, laying a foundation for more comprehensive document analysis.
- Integration of a multimodal relational extraction model that utilizes the output of the OCR process. This model is designed to:

$$Multimodal(OCR(I)) \approx \{((K_i, V_i) \mid K_i, V_i \in OCR(I)\}$$

  It processes both the text and its physical location, B_i, within the document, enhancing entity identification and classification by understanding the spatial relationships and textual-visual interplay.
- Application and enhancement of advanced text classification models to further refine the accuracy of NER. These models are adept at:

$$Classification(Multimodal(OCR(I))) \rightarrow L$$

  categorizing the relational pairs derived from the multimodal model into precise entities, leveraging the rich context established by preceding steps.

This methodical approach underscores our dedication to advancing document processing technology. By intricately merging analysis of semantic content with an understanding of spatial layouts, we strive for a nuanced and precise interpretation of complex documents. Our holistic strategy not only aims to improve the accuracy of entity recognition but also to enhance the overall context for information extraction, thereby redefining the standards in document analysis.

Ultimately, this research aspires to contribute significantly to the field of information extraction technologies, facilitating the development of more sophisticated and efficient systems for document

analysis. Through this work, we aim to set new benchmarks for intelligent document processing, marking a pivotal step forward in the evolution of NER systems.

2 Related Work

Tesseract OCR-Roberta: Introduced by Zacharias et al. in 2020, this model combines the robust OCR capabilities of Tesseract with the advanced natural language processing power of RoBERTa. While this integration leverages the strengths of both systems, it primarily relies on text-only embeddings, which limits its ability to fully comprehend the spatial dynamics of document layouts.

PPOCR-LayoutXLM-Bert: Developed by Zheng, Lianchi, et al. in 2022, this approach enhances OCR performance by integrating LayoutXLM with BERT to process textual content within documents. Despite its advancements, the model often underperforms in handling diverse and complex document structures, illustrating a common limitation in adapting to various layout contexts.

Problem Statement: The primary issues with the current models include an over-reliance on text-only embeddings and inadequate performance across different document layouts. These shortcomings highlight the need for models that incorporate spatial context into text embeddings, which could significantly enhance the understanding of document structures.

Objective: Our research aims to address these challenges by:

- Adding spatial context to text embeddings to enhance accuracy and comprehension of document spatial relationships.
- Adapting our model to better understand and interpret different document structures, ensuring it is robust against the variability in document types.
- Ensuring consistent high performance across a variety of document formats, thereby reducing the dependency on layout simplicity and increasing the model's applicability in real-world scenarios.

This section sets the stage for introducing our novel approach that integrates advanced spatial recognition algorithms with traditional text embeddings, aiming to overcome the limitations identified in the existing systems.

3 Methodology

3.1 Definition

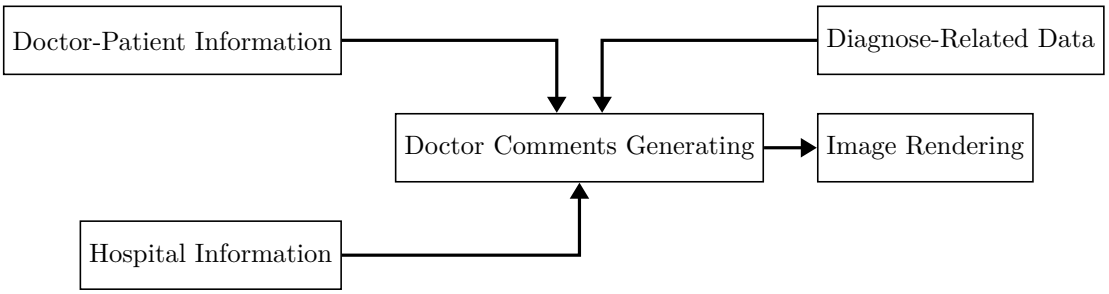
Evaluating a model's performance in document understanding tasks requires a comprehensive approach that considers various critical aspects: bounding box ‘s accuracy, the accuracy of text transcription, the accuracy of label assignment. we outline how each of these components is assessed to ensure a holistic evaluation of the model's capabilities:

1. **Bounding Box Detection:** The evaluation of bounding box detection focuses on how accurately the model identifies and locates each textual element within the document. Metrics such as Intersection over Union (IoU) are commonly used to measure the overlap between the predicted bounding boxes and the ground truth, with higher scores indicating greater accuracy.

2. **Text Regconition:** Text Regconition accuracy is assessed by comparing the model's output text against the actual text within the bounding boxes. Evaluation metrics such as character error rate (CER) or word error rate (WER) are employed to quantify the transcription accuracy, where lower rates indicate better performance.

3. **Label Assignment:** The correctness of label assignment is evaluated by examining how accurately the model classifies each text element into predefined categories or labels. Precision, recall, and F1 score are standard metrics used in this evaluation, reflecting the model's accuracy in understanding the semantic significance of text elements and assigning them appropriate labels based on their context within the document.

3.2 Data Collection



3.2.1 Doctor-Patient Information

Key to our dataset creation is the use of a specialized tool designed for auto-generating essential personal-related information, introducing realistic variability into the data. This includes doctor names, patient names, addresses, ages, sexes, and identification IDs. Such diverse personal information ensures our dataset accurately reflects real-world variability.

3.2.2 Hospital Information

We expanded our dataset with detailed hospital information obtained through web scraping from the Taiwan Ministry of Health and Welfare's database [1]. This includes hospital names, addresses, and department names, significantly enhancing the dataset's complexity and relevance to healthcare settings.

3.2.3 Diagnose-Related Data

Our dataset incorporates the first 100 ICD-10 codes, representing a broad spectrum of medical conditions and diagnoses, extracted from the Taiwan Ministry of Health and Welfare database [1]. This selection encompasses diverse health issues, providing a solid basis for analyzing healthcare trends and outcomes.

3.2.4 Doctor Comments Generating

We used a prompting technique with GPT-4 [2], accessed via the OpenAI API, to generate doctor comments. These comments are contextually relevant and medically informed, designed to reflect the nuances and specificities of each diagnosis, thereby enriching our dataset with detailed medical diagnostic information.

3.2.5 Image Rendering

To simulate realistic hospital documents, we collected 19 common Taiwanese hospital document templates and annotated bounding boxes based on their positions. Gaussian and uniform noise was applied to the rendered images to mimic natural variation found in real-world documents. The annotations were formatted according to the XFUND dataset specifications [3], ensuring readiness for document understanding evaluations.

Fig. 2: Left: Generated Image using our technique. Right: Real Image of a hospital document.

3.3 OCR model

In this study, we employ the PPOCR-v3 [4] model as the image encoder

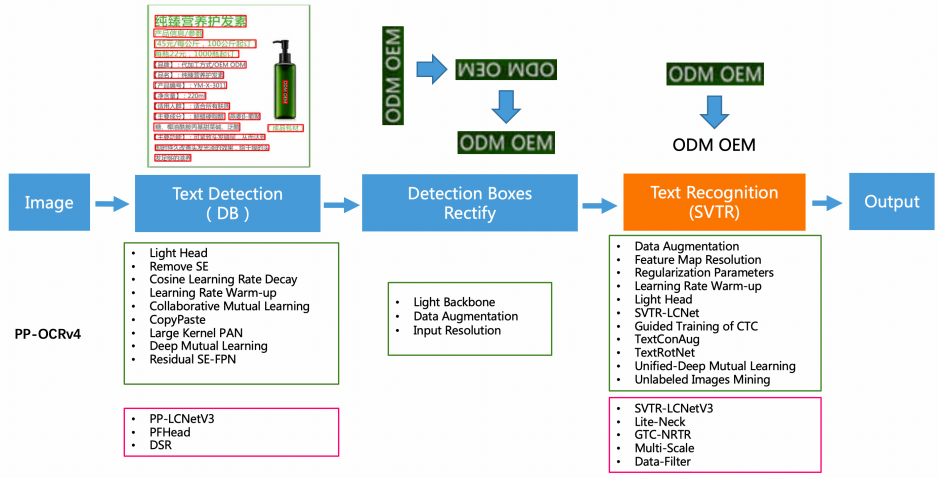


Fig. 3: Illustration of OCR System Enhancements and Their Impact

3.3.1 Text Detection

- **LCNetV3:** a more accurate backbone network, refinng feature extraction.
- **PFHead:** a parallel head branch fusion structure .
- **DSR (Dynamic Shrink Ratio):** Adaptive adjustment of shrink ratios during training .
- **CML (Cross Model Learning):** Application of KL divergence loss leveraging outputs from both Student and Teacher networks to fortify learning efficacy.

3.3.2 Text Recognition

- **SVTR_LCNetV3** A backbone network upgrade for heightened accuracy .
- **Lite-Neck:** Streamlining of the neck structure to facilitate efficient information flow.
- **GTC-NRTR:** Integration of a stable Attention guiding branch, enhancing focus on relevant features.
- **Multi-Scale Training:** Adoption of a versatile training strategy to accommodate varying text sizes and complexities.
- **DF (Data Filter):** utilizing an initial low-accuracy model to discard high-confidence redundant samples and a high-accuracy PP-OCRv3 model to remove low-confidence poor-quality samples, effectively reducing training data volume and time while enhancing accuracy

3.4 Key Information Extraction

Our research benefits from the enhancements in PP-StructureV2, particularly the visual-feature independent LayoutXLM (VI-LayoutXLM) model, which is optimized for quicker inference times while maintaining high accuracy levels.,The advent of LayoutLMv2 and LayoutXLM introduced visual backbone networks that significantly improved the models' capabilities to extract and merge visual features with textual embeddings, leading to a richer multi-modal input and enhanced contextual understanding. However, the visual feature extraction process, especially when utilizing ResNet x101 64x4d, adds substantial time overhead. To mitigate this, we have refined VI-LayoutXLM by excluding the visual feature extraction submodule.

3.4.1 Threshold-Based YX Sorting Algorithm

Typically, OCR results are sorted from top to bottom and left to right based on the detected text boxes' absolute YX coordinates. This method can lead to an order that does not naturally align with the expected reading order. To address this, the Threshold-Based YX Sorting Algorithm introduces a position offset threshold (th), enhancing the sorting mechanism. While the initial sorting from top to bottom is maintained, the introduction of th allows for an adjustment where, if the vertical distance between two text boxes is less than th, their order is then determined by their horizontal alignment. This adjustment ensures a sorting order that is more closely aligned with natural reading patterns.

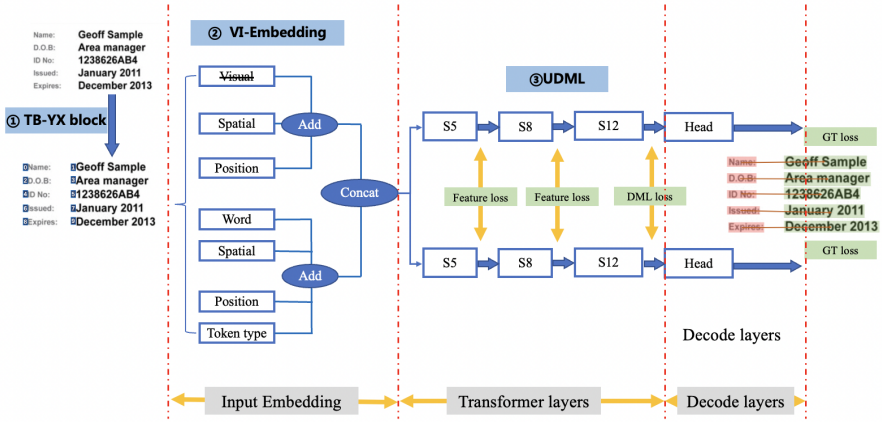
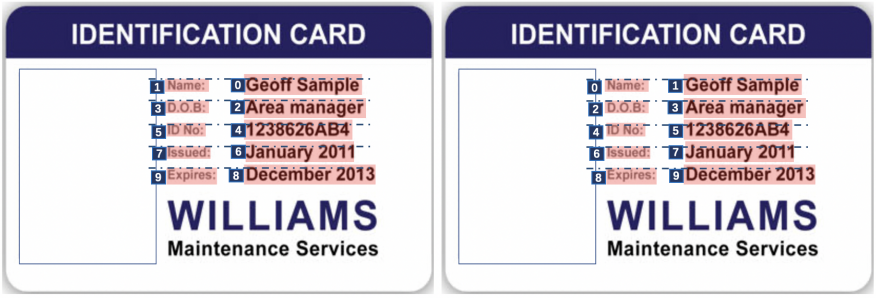


Fig. 4: XXXXX.



(a) YX sorting algorithm (a) TB-YX sorting algorithm

Fig. 5: Demonstration of the Threshold-Based YX Sorting Algorithm within VI-LayoutXLM, illustrating enhanced alignment with natural reading order.

3.5 Mutual Learning

We define a set of labels $Y = \{y_i\}_{i=1}^{N}$, where each label y_i corresponds to one of M possible classes. The probability of classifying an input x_i into class m by the neural network Θ_1 is given by the softmax function:

$$p_1^m(x_i) = \frac{\exp(z_1^m)}{\sum_{m=1}^{M} \exp(z_1^m)},$$

where z_1^m are the logits output from the network Θ_1 for each class m.

The network Θ_1 is trained using the cross-entropy loss, which quantifies the discrepancy between the predicted probabilities and the ground truth labels:

$$L_{C1} = -\sum_{i=1}^{N}\sum_{m=1}^{M} I(y_i, m)\log(p_1^m(x_i)),$$

with the indicator function $I(y_i, m)$ defined as:

$$I(y_i, m) = \begin{cases} 1 & \text{if } y_i = m \\ 0 & \text{if } y_i \neq m \end{cases}.$$

To enhance the generalization of Θ_1, a second network Θ_2 is utilized. Θ_2 also provides posterior probabilities p_2, and the alignment between the predictions of Θ_1 and Θ_2 is measured by the Kullback-Leibler (KL) Divergence:

$$D_{KL}(p_2 \parallel p_1) = \sum_{i=1}^{N}\sum_{m=1}^{M} p_2^m(x_i)\log\frac{p_2^m(x_i)}{p_1^m(x_i)}.$$

The overall loss function for Θ_1 integrates the cross-entropy loss with the KL divergence to penalize

prediction discrepancies between the two networks:

$$L_{\Theta 1} = L_{C1} + D_{KL}(p_2 \parallel p_1).$$

Similarly, Θ_2 is trained with a loss function that combines its own cross-entropy loss with the KL divergence from its predictions to those of Θ_1:

$$L_{\Theta 2} = L_{C2} + D_{KL}(p_1 \parallel p_2).$$

This dual training approach, where each network not only aims to correctly predict the labels but also to align its predictions with its peer, facilitates a deep mutual learning process, enhancing the generalization performance of both networks.

3.6 Relational Extraction Model

Building upon the work of Yang et al. [5], we adapt transformer-based architectures to classify relationship types in key-value pairs, framing it as a text classification challenge. Our model processes each pair by embedding and tokenizing the keys and values, utilizing special tokens ($[S1], [E1], [S2], [E2]$) to emphasize entity boundaries, thereby enabling precise classification based on the contextual relationship between the pair's components.

We explore four representation strategies to enhance classification accuracy:

1. Utilizing only the [CLS] token's embedding.
2. Merging the [CLS] token's embedding with that of entity start markers ($[S1], [S2]$).
3. Combining the [CLS] token's embedding with all entity marker embeddings.
4. Exclusively focusing on entity start marker embeddings.

Figure 3 illustrates our model's architecture, specifically designed to identify and classify relational types from structured key-value data.

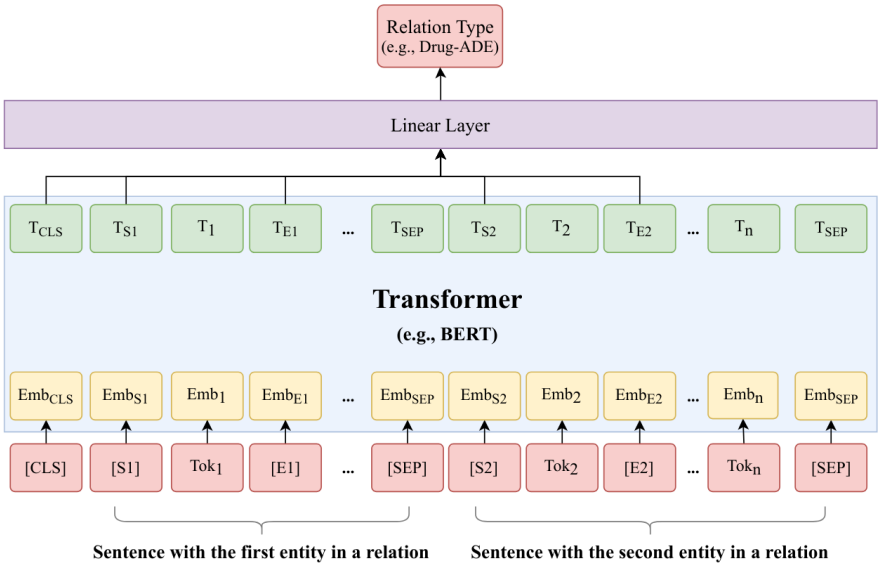


Fig. 6: Transformer model architecture for relational extraction.

This approach leverages the versatility of transformer models for the nuanced task of relational extraction, advancing the application of text classification methods to structured key-value pair analysis.

4 Experiment

4.1 Experiment Setup

4.1.1 Dataset

We generated 5,600 training data samples from the first 100 ICD-10 codes and collected 66 images from the internet for validation. All datasets are in Traditional Chinese and have been annotated across six categories: doctor's name, hospital name, hospital address, diagnosis, doctor comment, and department name. This specific language setting is crucial for ensuring the models trained are well-adapted to the linguistic nuances specific to Traditional Chinese medical documents.

Category	Training Data	Validation Data
Number of Images	5600	66
Templates	17	35
OCR annotation	3400	255

Table 1: Data Distribution

4.1.2 OCR Text Recognition

For OCR modeling we employed PPOCRv3 models utilizing the PaddleOCR lbirary [4]. This choice was motivated by its robust performance in recognizing textual content from images, including sentence-based text alongside their corresponding coordinate information.

4.1.3 Relation Extraction

In developing our relational extraction (RE) models, we built upon the advanced transformer architectures provided by the HuggingFace Transformers library [6] and PyTorch [7]. Initially, we engaged with transformer models pre-trained on general English text available in the HuggingFace repository as a baseline. To refine our approach for the medical domain, we selected specialized models, namely Bio-BERT [8] and Clinical-Longformer [9], both renowned for their pre-training on the English MIMIC-III dataset [10], a rich source of clinical narratives.

Despite the linguistic differences between our Traditional Chinese dataset and the English MIMIC-III corpus, both datasets are unified by their intrinsic clinical domain knowledge. This shared medical context underpins our decision to fine-tune Bio-BERT and Clinical-Longformer with our dataset. The fine-tuning process was tailored to bridge the language gap while capitalizing on the domain-specific insights these models have acquired through their initial pre-training. This involved adapting the models to better capture the nuances of clinical information presented in Traditional Chinese, leveraging their pre-existing understanding of clinical contexts derived from MIMIC-III.

Our fine-tuning regimen was structured around a comprehensive five-fold cross-validation strategy, aimed at optimizing hyperparameters such as training epochs and batch sizes, against the backdrop of a consistent learning rate of $1e-5$ and a fixed random seed of 13. Selecting the optimal model configuration was guided by the highest micro-averaged strict F1-scores obtained during cross-validation.

Integrating and fine-tuning Bio-BERT and Clinical-Longformer for our Traditional Chinese dataset represents a strategic pivot towards leveraging sophisticated AI technologies to improve relation extraction in clinical texts across languages. This methodology not only benefits from the domain-specific pre-training of these advanced transformer models on the MIMIC-III corpus but also innovatively adapts them to manage the complexities and specificities of medical narratives in a different language, thereby enhancing their utility and accuracy in extracting meaningful relationships from clinical documents.

4.2 Results

4.2.1 OCR Model Performance

Our OCR model was evaluated in two main areas: text detection and text recognition. The performance was assessed using standard metrics loss and accuracy for both tasks. The following tables summarize the results:

4.2.1.1 Text Detection and Recognition Performance

Task	Metric	Training Set	Evaluation Set
Text Detection	Loss	4.3	5.6
Text Recognition	Loss	0.03	0.07
Text Recognition	Acc	0.97	0.93

Table 2: Text detection and recognition performance metrics .

The results affirm that while the OCR model performs robustly on training data, real-world applications pose additional challenges, especially in text recognition. Ongoing efforts to refine data generation processes and enhance model training will focus on bridging the gap between synthetic training conditions and the variability encountered in real-world scenarios.

4.2.2 Key Information Extraction Model

The combined performance of VI-LayoutXLM and LayoutXLM in both semantic entity recognition (SER) and relation extraction (RE) tasks is summarized in the following table:

Model	SER				RE			
	Training		Validation		Training		Validation	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
LayoutXLM	0.98	0.99	0.85	0.9	0.97	0.98	0.88	0.85
Vi-LayoutXLM	0.97	0.97	0.92	0.88	0.95	0.97	0.9	0.88

Table 3: Detailed Comparative Performance of VI-LayoutXLM and LayoutXLM in Semantic Entity Recognition (SER) and Relation Extraction (RE) on Training and Validation Data

4.2.3 Relational Extraction Model

Table 4 presents the evaluation results for the Bio-Longformer and Bio-Transformer models in terms of precision, recall, and F1 score, showcasing their performance in the relational extraction task.

Model	Precision	Recall	F1 Score
Bio-Longformer	0.9	0.98	0.98
Bio-Transformer	0.949	0.948	0.948

Table 4: Evaluation of Relational Extraction Models

These results underscore the effectiveness of the models in accurately extracting relational data from clinical narratives. The Bio-Longformer demonstrates exceptional performance with perfect scores across all metrics, indicating its superior ability to understand and classify complex clinical relationships. The Bio-Transformer also shows high precision, recall, and F1 score, affirming its robustness in handling relational extraction tasks. The discrepancy in performance metrics reflects the inherent differences in model architecture and training, highlighting the Bio-Longformer's advantage in this specific application.

4.2.4 End to End System

The overall system evaluation assesses the integrated performance of different OCR and language processing models on six key categories: doctor name, diagnose, doctor comment, department name, hospital name, and hospital address. The models compared include:

- PPOCRv3-LayoutXLM (SER Only)
- TesseractOCR-Roberta
- PPOCRv3-LayoutXLM-Bert
- PPOCRv3-Vi-LayoutXLM-Longformer (Our Approach)

Category	**LayoutXLM**	**Roberta**	**LayoutXLM-Bert**	**ViLayoutXLM-Longformer**
Doctor Name	0.53	0.22	0.80	0.88
Diagnose	0.95	0.45	0.95	0.99
Doctor Comment	0.75	0.5	0.82	0.95
Department Name	0.88	0.45	0.90	0.90
Hospital Name	0.9	0.72	0.8	0.9
Hospital Address	0.6	0.61	0.85	0.93

Table 5: Comparative Performance on Key Categories

5 Limitations and Future Work

Our study faces several limitations that could affect its scope and applicability. Notably, the absence of open-source clinical datasets in Traditional Chinese limited our ability to comprehensively represent the spectrum of medical conditions. Additionally, the sensitive nature of human and hospital-related information posed significant challenges in data collection, impacting the depth of personal and institutional data included in our dataset. These limitations underscore the need for cautious interpretation of our findings and suggest directions for future research to overcome these challenges.

In many real-world scenarios, critical information does not always present itself in structured formats such as tables or graphs. These limitations highlight the necessity for further research and development to enhance the model's applicability and robustness across a broader range of data types and sources. Future efforts will focus on developing methods to integrate unstructured data effectively and exploring innovative approaches to simulate or anonymize sensitive information without compromising its utility for KIE tasks.

6 Conclusion

Our study reveals that transformer models, when applied to the relational extraction task within Key Information Extraction (KIE) systems, achieve higher accuracy compared to traditional approaches. This success is attributed to the transformer's inherent capability to analyze and predict relationships within key-value pairs, focusing precisely on the structured data typical of KIE models. However, the challenges of accessing comprehensive real-world datasets and the model's performance on unstructured information underscore the need for ongoing refinement and adaptation of these models. Future research is poised to address these limitations, aiming to expand the transformer's utility and efficacy in extracting and classifying relationships across a wider array of data formats and contexts, thereby enhancing the robustness and applicability of KIE systems in diverse medical informatics applications.