EMNLP 2024

Language, OCR, Form Independent (LOFI) pipeline for Industrial Document Information Extraction

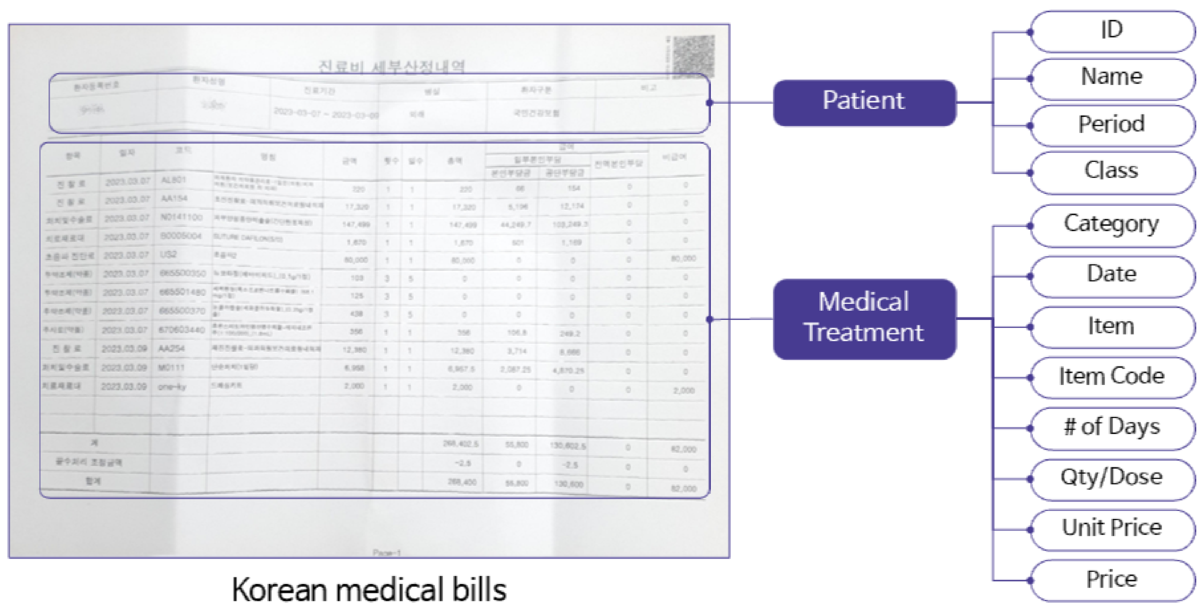
Chang Oh Yoon1,+, Wonbeen Lee1,+, Seokhwan Jang1,+,
Kyuwon Choi2,+, Minsung Jung2,+ ,
Daewoo Choi3,*

+AgileSoDA, *Hankuk University of Foreign Studies

Introduction

Many industries handle complex documents known as Visually Rich Documents (VRDs), containing text, tables, and figures. In real-world industry scenarios involving VRDs, we should consider a process of Semantic Entity Recognition (SER) to automate workflows.

For example, in insurance claims processing, patient information and diagnostic details need to be extracted from medical reports submitted by customers. Additionally, in accounting and tax filing processes, purchase information should be extracted from receipts or other tax documents.



Korean medical bills

Motivation

1 — Low Resource Language

- There are limited VRD datasets available for Low-Resource Languages.
- No pre-trained models exist for these languages.
- This scarcity hinders the creation of advanced language models.

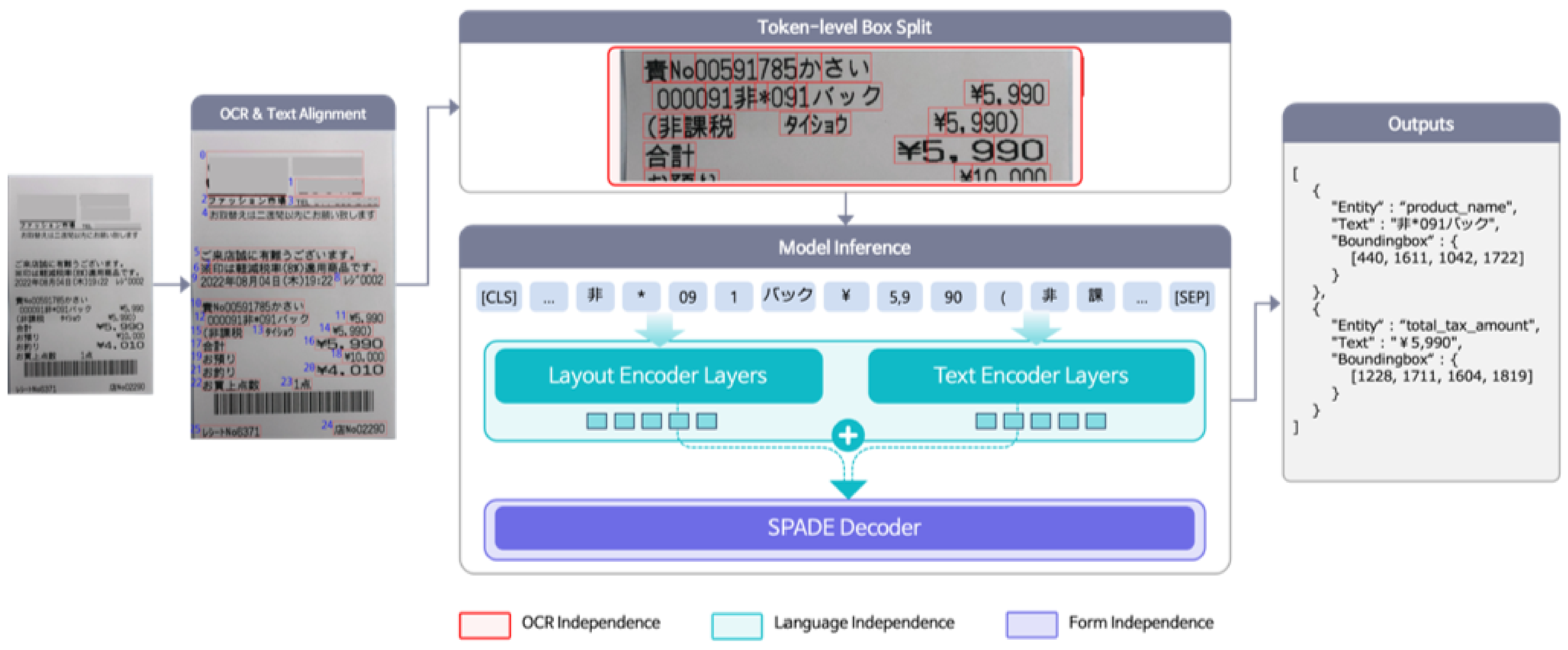
2 — OCR Dependency

- SER has limitations due to OCR engine output.
- OCR results are typically at the word level, not entity level.
- Additional processing (splitting or combining) may be needed for accurate semantic entities.

3 — Form Diversity

- Industry documents pose challenges for information extraction due to custom formats.
- Even standardized forms have variations in formatting, such as custom medical report templates.
- Image distortions or rotations can alter a document's structure and further complicate extraction.

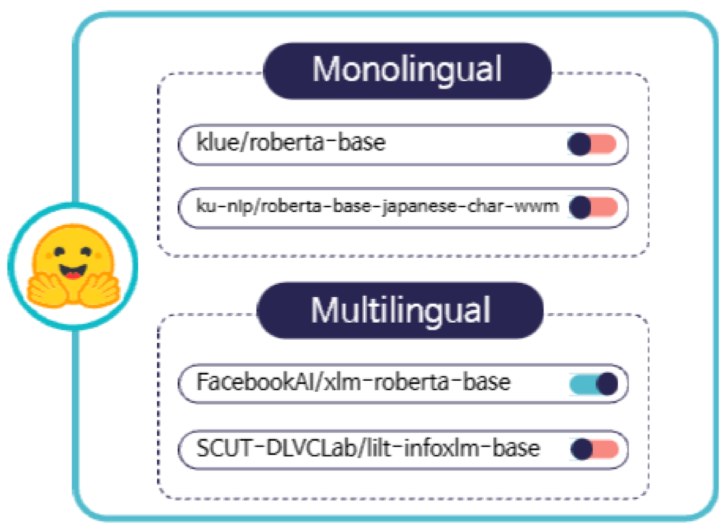
LOFI pipeline



An example of LOFI Pipeline for a Japanese receipt. Language independence is solved using Layout-independent Language Transformer (LiLT) as the backbone model as shown as a teal box. OCR independence is solved using token-level box split algorithm as shown as a red box. Form independence is solved using SPADE decoder as shown as a purple box.

To outline LOFI pipeline process:

1. *OCR and text alignment.* Our own OCR engine generates text and bounding box data from document images. Then, to preprocess 1D positional information, the results are sequentially arranged from top-left to bottom-right.
2. *Token-level box split.* Our own algorithm is applied to the sorted text and bounding boxes, to preprocess 2D positional information.
3. *Model inference.* The (token, token box) pairs are put into LiLT for sequence output generation. The SPADE decoder processes this output to produce ITC and STC results.
4. *Outputs.* The results are combined to generate the final SER output.

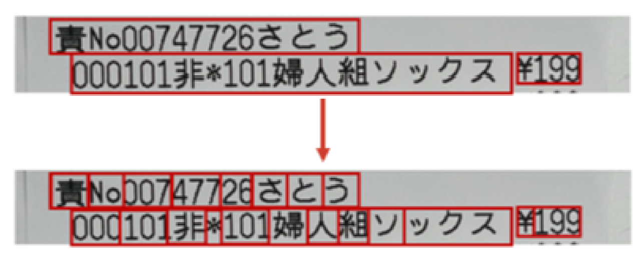


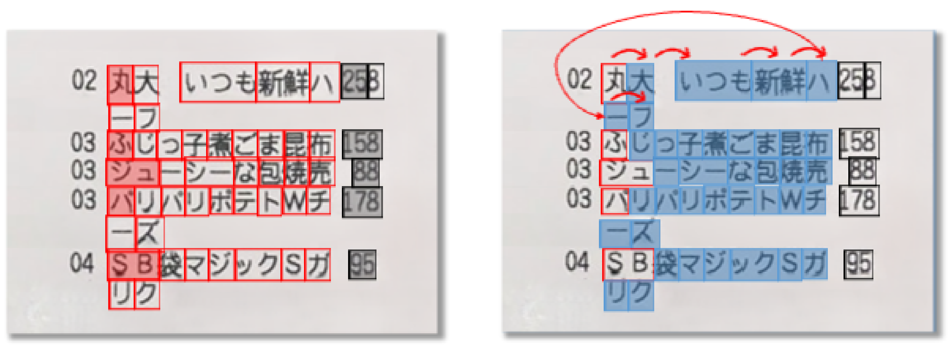
Language Independence

- Language models are paired with tokenizers, and Pretrained Language Models (PLMs) for specific languages typically use data predominantly in that language for tokenizer training.
- This ensures that tokens are structured to suit the characteristics of the language.
- LiLT utilizes a model structure that can adapt to the PLM corresponding to the language of the target document, enabling customized token configurations for Low-Resource Languages (LRL).
- Language-specific models tokenize sentences into more contextually relevant tokens compared to multilingual models, which may be less optimal for single-language tasks and could suffer from parameter inefficiencies.

OCR Independence

- The text and layout obtained through the OCR engine can have different ranges (character-level, word-level, line-level) depending on the linguistic and structural complexities of documents.
- This different range of bounding boxes results can lead to performance degradation in the SER model.
- We use the token-level box split algorithm to make the layout at the same level with any OCR engine.
- Tokens are proportionally assigned to bounding boxes based on character types, enabling a uniform split of the input into tokens with corresponding boxes.





Initial token classification *Subsequent token classification*

Form Independence

- In real-world scenarios, the numerous types of business documents used have diverse forms, limiting the ability to determine an appropriate reading order.
- The SPADE decoder operates robustly even with the incorrect order information by using the Initial Token Classification (ITC) and Subsequent Token Classification (STC) layer of the SPADE decoder.

Settings

Dataset

Dataset	Lang	Type	# of Entity	Train	Valid	Test
Medical Bills	Ko	Forms	68	829	98	-
Receipts	Ja	Receipts	15	990	110	-
FUNSD	En	Forms	3	149	50	-
CORD	En	Receipts	30	800	100	100

- Korean medical bills contain diverse medical and financial information from various Korean hospitals, including detailed patient records, treatment specifics, complex pricing tables, and hospital details.
- Japanese receipts contain information about the store name, expenditure details, taxes, etc, also in various types including mobile photos.

Model

Name	Pretrained Language Model
LOFI-en	Roberta-base (SCUT-DLVCLab)
LOFI-ko	Roberta-base (KLUE)
LOFI-ja	Roberta-base (Ku-NLP)
LOFI-mul†	InfoXLM-base (Microsoft)
LOFI-mul‡	XLMRoBERTa-base (FacebookAI)

- These are the PLM models used in the pipeline experiments.
- Each of these models is combined with LiLT's Layout encoder.

Experiment results

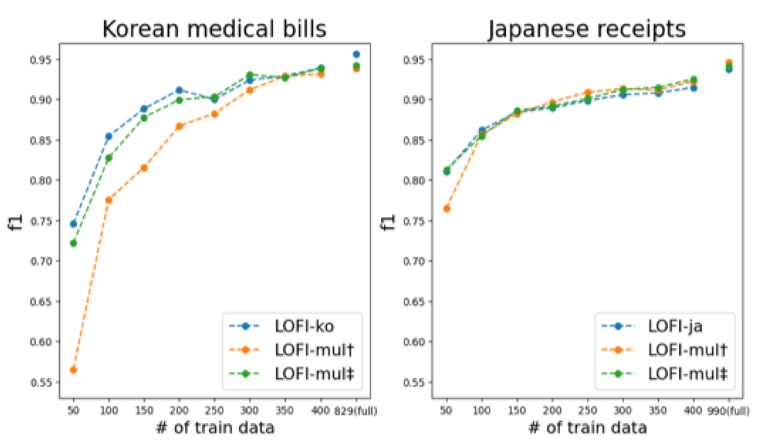
1. LRL business documents

Name	Language	Encoder	Parameters	Modality	Image Embedding	Korean medical bills	Japanese receipts
LayoutXLM°	Multi	$\text{LayoutXLM}_{\text{BASE}}$	369 M	T + L + I	ResNeXt101-FPN	95.58%	94.35%
LOFI-mul†	Multi	$\text{InfoXLM}_{\text{BASE}}$ + lilt-only-base	284 M	T + L	None	93.81&	**94.60%**
LOFI-mul‡	Multi	$\text{XLMRoBERTa}_{\text{BASE}}$ + lilt-only-base	284 M	T + L	None	94.24&	94.10%
LOFI-ko	Ko	$\text{RoBERTa}_{\text{BASE}}$ + lilt-only-base	116 M	T + L	None	**95.64%**	-
LOFI-ja	Ja	$\text{RoBERTa}_{\text{BASE}}$ + lilt-only-base	106 M	T + L	None	-	93.78%

2. Open datasets

Name	Parameters	Modality	Image Embedding	FUNSD	CORD
LayoutLM	160 M	T + L	ResNet-101 (fine-tune)	79.27 %	94.72 %
LayoutLMv2	200 M	T + L + I	ResNeXt101-FPN	82.76 %	94.95 %
LayoutLMv3	133 M	T + L + I	Linear	79.38 %	**96.80 %**
BROS	110 M	T + L	None	**83.05 %**	95.73 %
LOFI-en	131 M	T + L	None	78.99 %	96.39 %

3. Number of training data



We use the entity-level F1 score as the measure standard for both experiments.

1. LOFI-ko and LOFI-mul † demonstrate better F1 score with fewer parameters compared to LayoutXLM on Korean medical bills and Japanese receipts, respectively. This result highlight the effectiveness of LOFI for LRL documents, even in the absence of specific PLMs.
2. LOFI-en was similar to LayoutLMv3 on CORD (96.39%) but trailed BROS by 4% on FUNSD (78.99%). This reveals LOFI's need for ample finetuning data, evident in performance differences between CORD (800 documents) and FUNSD (149 documents).
3. The required number of training data may differ based on language, document structure, and characteristics. But, achieving satisfactory performance typically requires at least 300-400 documents. With fewer than 200 training documents, there is at least 5% performance difference compared to using the full training dataset.

Agile SoDA