Improving Information Extraction on Business Documents with Specific Pre-Training Tasks

Thibault Douzon, Stefan Duffner, Christophe Garcia, Jérémy Espinas

Improving Information Extraction on Business Documents with Specific Pre-Training Tasks

Thibault Douzon[1,2] (✉), Stefan Duffner[1], Christophe Garcia[1], and
Jérémy Espinas[2]

[1] INSA Lyon, LIRIS, Lyon, France
firstname.lastname@insa-lyon.fr
[2] Esker, Lyon, France
firstname.lastname@esker.com

Abstract. Transformer-based Language Models are widely used in Natural Language Processing related tasks. Thanks to their pre-training, they have been successfully adapted to Information Extraction in business documents. However, most pre-training tasks proposed in the literature for business documents are too generic and not sufficient to learn more complex structures. In this paper, we use LayoutLM, a language model pre-trained on a collection of business documents, and introduce two new pre-training tasks that further improve its capacity to extract relevant information. The first is aimed at better understanding the complex layout of documents, and the second focuses on numeric values and their order of magnitude. These tasks force the model to learn better-contextualized representations of the scanned documents. We further introduce a new post-processing algorithm to decode `BIESO` tags in Information Extraction that performs better with complex entities. Our method significantly improves extraction performance on both public (from 93.88 to 95.50 F1 score) and private (from 84.35 to 84.84 F1 score) datasets composed of expense receipts, invoices, and purchase orders.

Keywords: Business Documents · Document Understanding · Information Extraction · Pre-Training · `BIESO` Decoding · Transformer

1 Introduction

Business documents are paper-sized files containing useful information about interactions between companies. They may take the form of invoices, purchase orders, various reports, and agreements. The exact layout of a document depends on the issuer, but the contained information is conventionally structured. For example invoices and purchase orders share the same header, table, footer structure that almost all issuers have adopted. Because such documents trace every transaction made by companies, they are the key to business process automation. With the emergence of modern resource planning systems, accurate Information Extraction (IE) has become one of the core problems of Document Intelligence.

Initially, information extraction was done by human operators, but software solutions have been developed since the early days of document analysis to tackle the problem. Their intent was to ease the work of human operators with hard-coded extraction rules. Unfortunately, these rules needed to be adapted for each and every layout of documents. This limitation has led to the rise of Machine Learning (ML) models for automatic document IE.

First ML approaches relied on Optical Character Recognition (OCR) systems to provide the textual content of the document. This transition from image to text allowed for standard Natural Language Processing (NLP) methods to be applied by adopting a Sequence Labeling problem. The amount of labeled data necessary to train accurate NLP models has always been a problem. Business documents are inherently private which strongly limits the quantity of publicly available data. Thus, only companies selling business process automation software are able to collect larger amounts of such data. Moreover, they often rely on their customer to implicitly label the documents.

Most recent proposals often include a pre-training step, where the model is trained on a pretext task. Those pretext tasks are self-supervised problems that teach the model many useful “skills” for manipulating the data. Usually, these tasks are as broad as possible, teaching the model common sense about language, grammar, and global structure.

In this work, we focus on LayoutLM [31], a pre-trained Transformer that is specialized in business documents. As shown in Fig. 1, it reuses the same Transformer layer with multi-head attention with the addition of a 2D positional encoding. Its larger version achieved state-of-the-art performance in both document classification and information extraction. However, the required hardware to train it can be repelling. In this paper, we propose new pre-training tasks specific to business documents that will provide additional skills to the model. We also propose a new decoding post-processing algorithm that prevents many

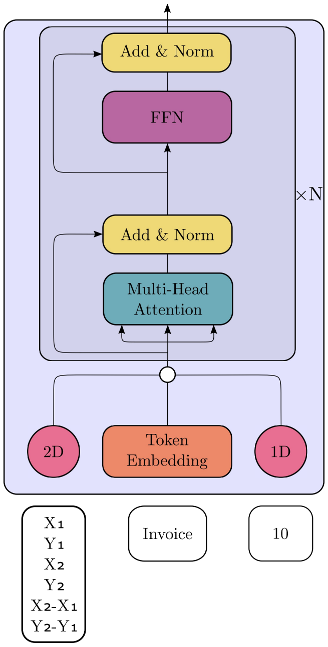


Fig. 1: LayoutLM architecture.
Token embeddings are enriched with 1D positional encoding and 2D spatial encoding specific to this architecture. The number of blocks N varies from 12, for the base model, to 24, for the large one.

errors made by the model due to ambiguities. Combined, our contributions[3] allow for the base LayoutLM model to perform on par with the large version.

2 Related Work

2.1 Information Extraction

Rule-based approaches [10] have been supplanted by Deep Learning models in the last decade. Document IE first capitalized on the state of the art in Named Entity Recognition for NLP [8]. Recurrent Neural Networks with Long-Short Term Memories were first used to encode documents at a word level [16; 23], allowing a simple classifier to predict each word's associated label. Instead of a softmax and cross-entropy loss, a Conditional Random Field [25] model has been used in addition to `BIESO` tags. Other architectures have also been proposed to better adapt to the specificity of the document. For example, graphs [13; 12; 33; 3] and convolutions over a grid [7; 1; 11] constrained the model based on the words' positional information. Because most architectures relied on textual representations, they benefited from pre-trained word embeddings like Word2Vec [14] or GloVe [17].

With the emergence of Transformers [26] and text encoders like BERT [2], attention-based document analysis models [31; 5; 30] evolved quickly, which resulted in a large improvement of state-of-the-art performance. In line with [7] which included both textual and visual representations, multi-modal Transformers [30; 11] superseded conventional textual models.

In parallel to the rise of Transformers, end-to-end IE models tried to reduce the labeling cost. First using RNNs with attention layers [15; 22], then shifting to Transformers [19]. Adopting at the same time the Question Answering [4] (QA) format, instead of the usual Sequence Labeling, provided more flexibility on the predicted labels.

2.2 Pre-Training

Semi-supervised training and pre-trained models were popularised in NLP with enriched word embeddings [18; 14; 17]. With the emergence of Transformers, large pre-trained models have been proposed [29]. Thanks to their pre-training, they can efficiently adapt to various tasks [28; 27] and data types. In general, these models are pre-trained on large unlabeled datasets in a self-supervised manner. This self-supervision removes parts of the burden of data labeling [24] and leverages the huge quantities of available data.

A wide variety of pre-training tasks have been proposed. General-purpose tasks aiming at learning the language and grammar were used first. Auto-regressive tasks [20] and Masked Language Modeling [2] are still frequently used in new pre-trained models as they have proven to be effective in most situations. In addition to incremental improvements [21; 32], some new pre-training tasks were designed to align representations of multi-modal inputs [30; 19].

[3] Code available here: `https://github.com/thibaultdouzon/business-document-pre-training`

3 Models

3.1 Architecture

We used the well-established LayoutLM architecture [31] which itself is based on BERT Transformer [2]. More specifically, we chose the base model[4] with 12 layers and 512 dimensions for token embeddings. This model is computationally much more efficient compared to the larger version while still giving very good performance.

Transformer models work on tokens that are in between characters and words. LayoutLM and BERT both use the WordPiece algorithm. We use the same tokenizer as LayoutLM in order to compare our performance with the base LayoutLM model. It uses a vocabulary size of 30000, and we limit the sequence length to 512 tokens, including the special tokens `[CLS]` and `[SEP]`. This limitation due to GPU memory consumption of self-attention operations often forces us to cut documents in multiple pieces of 512 tokens and process them separately.

Contrary to RNNs, all positions in the sequence are equivalent in a Transformer model. To provide information about position inside the sequence, a linear positional encoding [26] is added for each token. Then LayoutLM adapted this positional encoding to a 2D version that can represent the positions of words on a page.

For both pre-training tasks and fine-tuning, we use a simple dense layer to map each token's final internal representation to the dimension of the prediction space. A softmax layer is applied to produce the final model confidence scores. For training, the cross-entropy loss is used on the model confidence scores.

3.2 ConfOpt Post-Processing

We model the Information Extraction task as sequence tagging on tokens. Predictions are done at the token level and then aggregated by a lightweight post-processing step to give the model's final prediction. In all experiments, we use `BIESO` tagging. That is, each field to extract is composed of a sequence of target tags of the following types: `B` for the beginning of the entity, `I` for inside, `E` for its end, or otherwise `S` for a single token entity. `O` is used for any token that is outside any target label. `BIESO` is widely used in IE as it provides structure to the target sequence that helps the model.

Instead of the trivial post-processing which consists of simply following the maximum confidence of the model, we decided to decode a model's prediction by solving a basic optimization problem. We will refer to this method as ConfOpt in the remaining of the paper. The predicted sequence for a target label is the sequence that maximizes model confidence over the whole input sequence. There is a constraint to decode a prediction: it must match the following regular pattern: `(BI*E)` | `S` where `*` denotes zero or many occurrences and | denotes an alternative.

[4] Pre-trained weights available here: `https://huggingface.co/microsoft/layoutlm-base-uncased`

This optimisation problem can be solved with a dynamic programming approach. The model's predictions for one target label can represented as a $4 \times N$ dimensional matrix where N is the sequence length and 4 comes from the 4 tags `B,I,E,S`. By noting $C_{T,0}$ the model's confidence in `T` tag at position 0 and $P_{T,i}$ the best prediction confidence ending at token i with tag `T`, the objective is to determine $S = \max_{\substack{0 \leq i < N \\ T \in \{E,S\}}} P_{T,i}$ where

$$P_{B,i} = C_{B,i} \ ; \ P_{I,i} = C_{I,i} + \max \begin{cases} P_{B,i-1} \\ P_{I,i-1} \end{cases}$$

$$P_{S,i} = C_{S,i} \ ; \ P_{E,i} = C_{E,i} + \max \begin{cases} P_{B,i-1} \\ P_{I,i-1} \end{cases}$$

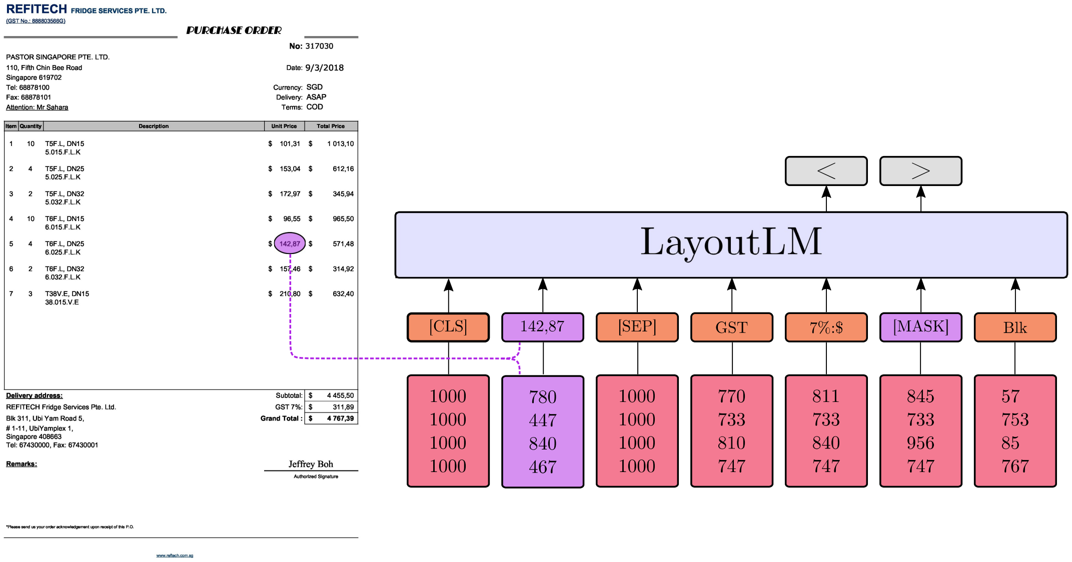
One drawback of this post-processing is dealing with no prediction and non-unique predictions. It can be solved with an empirically determined threshold below which no predictions are made. Though in this paper this is not further studied because fields are mandatory in a document and always unique.

4 Pre-training

Transformer models provide great performance when first pre-trained on pretext tasks on very large unlabelled datasets. This pre-training is most of the time done in a self-supervised manner in order to avoid the labeling cost. LayoutLM uses Masked Visual-Language Modeling [31] which is adapted from BERT's Masked Language Modeling [2]. It teaches the model how text and documents are formed at a token level. In practice, at each training step, 15% of the tokens are randomly chosen and replaced by either a `[MASK]` token, a random token, or not replaced at all. The model tries to guess which token is the most probable right replacement at those positions.

For all pre-training tasks when a document is too long to be processed at once, we randomly select a continuous span of words of maximum size and provide it to the model instead. We expect the model to learn useful features on various parts of documents thanks to the long training. For very short documents, the input is padded to the maximum size.

We introduce two new specific pre-training tasks in addition to Masked Visual-Language Modeling (MVLM). The first one, Numeric Ordering teaches the model how to compare and order numbers. The second one, Layout Inclusion focuses on words in the 2D plane and their relative positioning. We chose to avoid regression tasks, even though their implementation would have been simpler. For example, simply removing the 2D positioning of some tokens, and asking the model to predict tokens' position is an alternative to what we propose. But this does not behave well for a token that could appear either at the top or the bottom of the document: the model would learn its mean position – the middle – where the token would never appear. In the following, we will describe the two pre-training tasks in detail.



REFITECH FRIDGE SERVICES PTE. LTD.
(GST No.: 888803566G)

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68878100
Fax: 68878101
Attention: Mr Sahara

Date: 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	T5F.L, DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	T5F.L, DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	T5F.L, DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	T6F.L, DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	T6F.L, DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	T6F.L, DN32 6.032.F.L.K	$ 157,46	$ 314,92
7	3	T38V.E, DN15 38.015.V.E	$ 210,80	$ 632,40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yam Road 5,
1-11, UbiYamplex 1,
Singapore 408663
Tel: 67430000, Fax: 67430001

Subtotal:	$ 4 455,50
GST 7%:	$ 311,89
Grand Total :	$ 4 767,39

Remarks:

Jeffrey Boh
Authorized Signature

*Please send us your order acknowledgement upon receipt of this P.O.

www.reftech.com.sg

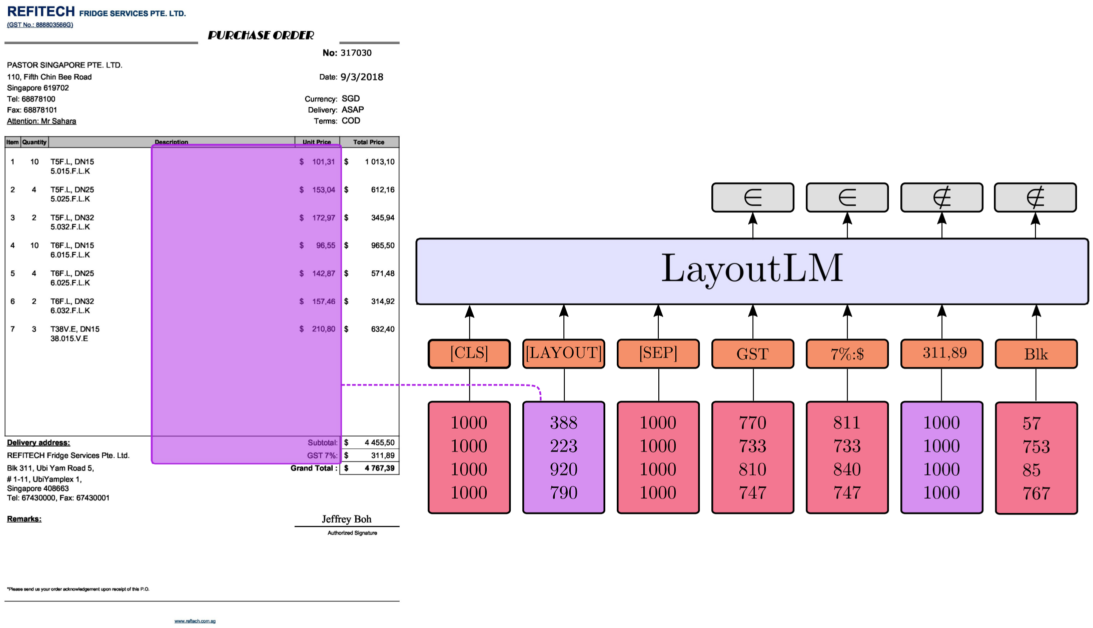
Fig. 2: A pre-training example with Numeric Ordering task. A random token containing a number is selected, then the target is to predict whether other numbers are smaller or bigger. Some random noise can be added by masking tokens' textual or spatial representations. Only a small part of the document's input is represented in this illustration.

4.1 Numeric Ordering Task

Numeric Ordering (NO) focuses on numeric figures in the document and their relative values. Contrary to MLM which only relies on self-supervised data, NO relies on a handcrafted number parser to find and parse all numbers that appear in a document. Because business documents are mostly made of decimal numbers written with digits, we ignore those written out in words. The numeric value of each token is determined by parsing beforehand each word in the document, looking for numbers and ignoring irrelevant characters.

As shown in Fig. 2, the model must predict for every numeric figure in the document if its parsed value is smaller, equal or greater than a randomly selected number among the document. The loss is only computed on tokens starting a new word, but tokens continuing a word are important to determine the value represented by a word.

We want the model not only to reason on the textual features, but also on the spatial context surrounding each figure in the document. Therefore, we randomly mask the textual representations of 15% of the numbers in the document and replace them with the `[MASK]` token as shown in Fig. 2. For the same reason, we also mask the spatial encoding of 15% of the numbers and make sure both text and position are not masked at the same time. All masked positions are replaced with (1000, 1000, 1000, 1000).



REFITECH FRIDGE SERVICES PTE. LTD.
(GST No.: 888803566G)

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68878100
Fax: 68878101
Attention: Mr Sahara

Date: 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	T5F.L, DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	T5F.L, DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	T5F.L, DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	T6F.L, DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	T6F.L, DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	T6F.L, DN32 6.032.F.L.K	$ 157,46	$ 314,92
7	3	T38V.E, DN15 38.015.V.E	$ 210,80	$ 632,40
			Subtotal:	$ 4 455,50
			GST 7%:	$ 311,89
			Grand Total :	**$ 4 767,39**

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yam Road 5,
1-11, UbiYamplex 1,
Singapore 408663
Tel: 67430000, Fax: 67430001

Remarks:

Jeffrey Boh
Authorized Signature

*Please send us your order acknowledgement upon receipt of this P.O.

www.reftech.com.sg

Fig. 3: A pre-training example with Layout Ordering task. Coordinates of the purple rectangle are drawn uniformly. Random noise is added by masking the 2D position of some tokens. Only a small part of the document is represented.

4.2 Layout Inclusion Task

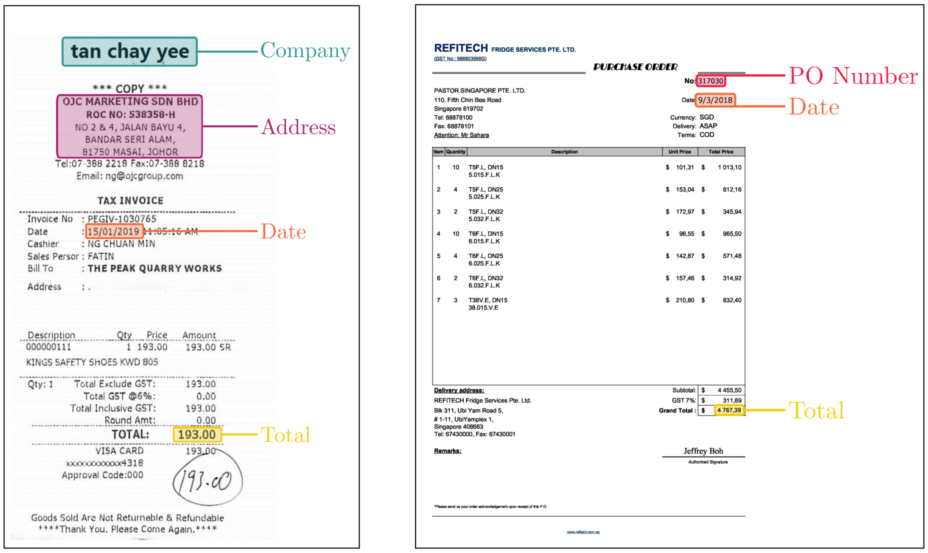
We introduce another pre-training task focusing on the 2D positional encoding, which we called Layout Inclusion (LI). Its purpose is to provide a better understanding of document layouts and complex structures. In fact, most business documents, including invoices, purchase orders, and expense receipts, contain tables where the meaning of tokens is mostly driven by their position relative to headers.

As shown in Fig. 3, Layout Inclusion is formatted like a question answering prompt: a question followed by the content of the document. The question is simply a special token `[LAYOUT]` positioned at random coordinates (x_1, y_1, x_2, y_2). The model must then classify every token in the document into 2 groups: either `inside` or `outside` of the question token. More precisely, the target answer is whether the middle point of a document token is inside or outside the rectangle described by the coordinates of the question.

Again, the objective is for the model to not only reason on the 2D positions of tokens but also use their textual embedding. In order to force the model to use both representations, we randomly replace 15% of documents token positions with (1000, 1000, 1000, 1000). In case of a random position replacement, the target value is still computed based on the real position of the token, and the model must make its prediction based on the token's text and the neighboring tokens using the classical 1D positional encoding.

5 Datasets

We used 2 different collections of documents to build 3 datasets for training and evaluation as described in the following. They all contain business documents: invoices and purchase orders for the private collection and expense receipts for the public one. The largest dataset used for pre-training isn't labeled, document samples with their target fields for the others datasets are shown in Fig. 4.



tan chay yee

*** COPY ***
OJC MARKETING SDN BHD
ROC NO: 538358-H
NO 2 & 4, JALAN BAYU 4,
BANDAR SERI ALAM,
81750 MASAI, JOHOR
Tel:07-388 2218 Fax:07-388 8218
Email: ng@ojcgroup.com

TAX INVOICE

Invoice No : PEGIV-1030765
Date : 15/01/2019 11:05:16 AM
Cashier : NG CHUAN MIN
Sales Persor : FATIN
Bill To : THE PEAK QUARRY WORKS
Address : .

Description	Qty	Price	Amount
000000111	1	193.00	193.00 SR
KINGS SAFETY SHOES KWD 805			

Qty: 1 Total Exclude GST: 193.00
Total GST @6%: 0.00
Total Inclusive GST: 193.00
Round Amt: 0.00
TOTAL: 193.00
VISA CARD 193.00
xxxxxxxxxxxx4318
Approval Code:000
193.00

Goods Sold Are Not Returnable & Refundable
****Thank You. Please Come Again.****

(a) Receipt from SROIE.

REFITECH FRIDGE SERVICES PTE. LTD.
(GST No.: 888803566G)

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68878100
Fax: 68878101
Attention: Mr Sahara

Date 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	T5F.L, DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	T5F.L, DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	T5F.L, DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	T6F.L, DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	T6F.L, DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	T6F.L, DN32 6.032.F.L.K	$ 157,46	$ 314,92
7	3	T38V.E, DN15 38.015.V.E	$ 210,80	$ 632,40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yam Road 5,
1-11, UbiYamplex 1,
Singapore 408663
Tel: 67430000, Fax: 67430001

Subtotal: $ 4 455,50
GST 7%: $ 311,89
Grand Total : $ 4 767,39

Remarks:

Jeffrey Boh
Authorized Signature

*Please send us your order acknowledgement upon receipt of this P.O.

www.reftech.com.sg

(b) Purchase Order from BDC-PO.

Fig. 4: A document sample for each training dataset annotated with the expected predictions. For BDC-PO, we replaced the document with a fictive one due to privacy reasons.

5.1 Business Documents Collection

The Business Documents Collection (BDC) is a large private dataset composed of 100k invoices and 300k purchase orders. Those real documents were submitted and processed on a commercial document automation solution in the last 3 years. It contains English-only documents divided into 70000 different issuers. All documents sharing the same issuer usually use the same information template. Therefore, we limited the maximum number of documents of the same issuer to 50. It is important to keep the number of similar layouts in the collection low and the variety of examples high. We used this collection for pre-training language models on business documents that are closer to our final objective than RVL-CDIP [9].

Textual and positional information have been extracted using a commercial OCR system. It achieves excellent accuracy on properly scanned documents and provides accurate word positions. We also use the provided read order to determine the order of tokens when feeding the network. This order determines the 1D positional encoding given to each token that complements the 2D positional encoding.

Because we only used this collection for pre-training models on self-supervised tasks, most documents do not have extraction labeling. Only a subset composed of purchase orders is labeled for the IE task.

5.2 Business Documents Collection – Purchase Orders

We selected a subset of the Business Documents Collection to build a labeled dataset of English purchase orders called BDC-PO. It contains almost 9000 different issuers split into training, validation, and test set. In order to not introduce bias for models pre-trained on the BDC, we removed from BDC all documents emitted by a supplier contained in the test set. This means that document layouts contained in the test set have never been seen before by the model at pre-training or training time.

Long purchase orders are rare but can sometimes be longer than 20 pages. If we wanted to train models and make predictions on such documents, we would have to evaluate the model on dozens of inputs for one document. Instead, we chose to limit documents to one page and crop the remaining. It only concerns roughly 25% of the dataset and sometimes impacts the prediction because labels are missing from the input.

The extraction task consists of 3 fields: document number, delivery date, and total amount. Those fields were chosen because they are mandatory for most customers and thus are well labeled at the word level by the end-user. We controlled the labeling quality at the issuer level and rejected from the dataset some issuers with undesirable labeling practices.

5.3 ICDAR 2019 – Scanned Receipts

We also trained and evaluated our model on the public Scanned Receipts OCR and Information Extraction [6] (SROIE) dataset that was published for ICDAR 2019. We focus on the third task which consists in extracting information from the documents. SROIE contains Malaysian receipts split into 626 train and 347 test documents. Unfortunately, we do not have control over the composition of the test set, and most of the test layouts also occur in the training set.

We used the OCR text provided with the dataset instead of using our own OCR system. As others have pointed out [31], it contains numerous little errors that negatively affect the final performance. For a fair comparison with the leaderboard, we manually fixed them such that the expected string appears in the input, at least. These fixes mostly concern addresses and company names. It almost exclusively involves fixing errors related to white-spaces and commas.

6 Experiments

All experiments were performed on a single machine equipped with two Nvidia RTX A6000 with 48Go of video memory each. This allowed us to boost the batch size up to 32 per device on a base transformer model. To further increase the batch size, we also aggregated 6 batches together before propagating the gradient for a total batch size of 192. We used the Adam optimizer with a learning rate of $1e-5$ and 200 linear warm-up steps as it improved our model's convergence. We used 1500 training steps for SROIE and 3000 steps for BDC-PO. Finally, we ran each fine-tuning 10 times in each setup to get a precise idea of the performance of the models and the variability of the results. For the different pre-training scenarios, we performed only two runs and the best model was kept.

6.1 Post-Processing

This first set of experiments aims at comparing the post-processing used to decode the sequence produced by the model. We want to determine whether our proposed ConfOpt algorithm is competitive with other decoding methods. We decided to use the LayoutLM base model and compare the proposed ConfOpt against two other decoding algorithms as shown in Table 1.

We named Ad-Hoc the basic decoding using the label with maximal confidence for each token. When decoding with this method, a `B` tag starts a new entity, a `I` tag continues the previous entity, a `E` closes the previous entity, and a `S` tag produces a new entity and closes it right away. Ad-Hoc and ConfOpt use the same model weights in this experiment as they do not introduce any trainable parameters.

The second decoding algorithm uses a Conditional Random Field (CRF) [25; 8] that processes LayoutLM's predictions. In this particular case, we did not use the classical cross-entropy loss but the score provided by the CRF layer. Because the CRF required specific training and did not optimize the same loss, its weights are different from the two other post-processing methods.

	Fine Tuning (F1 score)	
Post Processing	SROIE	BDC-PO
Ad-Hoc	93.88 ± 0.59	84.35 ± 0.12
CRF	94.01 ± 0.55	84.40 ± 0.16
ConfOpt	$\mathbf{94.94 \pm 0.38}$	$\mathbf{84.57 \pm 0.10}$

Table 1: Performance comparison on SROIE and BDC-PO between multiple post-processing algorithm. Score is computed on the exact match between the prediction and the target string.

We evaluated these algorithms on both SROIE and BDC-PO. The results in Table 1 show a tiny improvement using a CRF instead of the Ad-Hoc post-processing (0.13 and 0.05 F1 points) but those differences are always within one standard deviation range. We would need more evidence to conclude on the effect of adding a CRF layer for the post-processing.

On both datasets, using ConfOpt significantly increases performance (1.06 and 0.22 F1 points) compared to the Ad-Hoc post-processing, even though the model is strictly identical. In light of these results, we decided to use the ConfOpt for the next experiment.

6.2 Business Document-Specific Pre-training

We conducted another set of experiments in order to study the effects of the new business data-specific pre-training tasks on the model performance. At the same time, we controlled the performance gap obtained by pre-training with the basic MVLM task on the same new dataset. Both comparisons are insightful to decide whether it is useful to pre-train on clients' data and/or with data-specific pre-training tasks.

For the pre-training part, we always initialize the model's weights with the base version [31]. We pre-train models for 20 epochs on 80% of BDC. When using multiple pre-training tasks at the same time, we chose to provide batches of single tasks to the model. Gradient aggregation over multiple batches helps smoothing the update between different tasks. We pre-trained 2 models on the BDC, one with MVLM only and another with MVLM+NO+LI.

Pre Training			Accuracy per field		
Task(s)	Dataset	F1 Score	PO Number	Total	Date
MVLM	RVL-CDIP	84.57 ± 0.10	89.98	89.10	93.59
MVLM	BDC	84.77 ± 0.12	90.61	89.33	93.59
MVLM+NO+LI	BDC	$\mathbf{84.84 \pm 0.08}$	**90.71**	**89.36**	**93.83**

Table 2: Model performance when fine-tuning on BDC-PO

Architecture	Pre Training			Accuracy per field			
	Task(s)	Dataset	F1 Score	Company	Address	Total	Date
LayoutLM base *	MVLM	RVL-CDIP	94.94 ± 0.38	92.91	90.81	89.25	99.48
LayoutLM base *	MVLM	BDC	95.18 ± 0.23	**93.72**	91.00	89.48	**99.68**
LayoutLM base *	MVLM+NO+LI	BDC	$\mathbf{95.50 \pm 0.22}$	93.60	**91.41**	**90.89**	99.57
LayoutLM base [31]	MVLM	RVL-CDIP	94.38				
LayoutLM large [31]	MVLM	RVL-CDIP	95.24				
LayoutLMv2 large [30]	MVLM+TIA+TIM	RVL-CDIP	**97.81**				

Table 3: Model performance when fine-tuning on SROIE. Models name ending with a * are our contribution. The second part contains published scores of the original LayoutLM and LayoutLMv2 as a comparison.

We evaluated each pre-trained model on both datasets, the results are available in Table 2 for BDC-PO and Table 3 for SROIE . Each cell contains the means of 10 runs with different seeds and the standard deviation is provided for the F1 score. There are a few interesting things to notice.

The first important remark is the importance of the pre-training dataset. Pre-training on BDC significantly improves performance on both SROIE and

BDC-PO, even though the pretext training task is the same as what was used for LayoutLM. BDC is more homogeneous and focuses on invoices and purchase orders. Contrary to our expectations, we observe a greater improvement on SROIE than on BDC-PO (0.24 vs 0.2 F1 points). But the overall improvement by using BDC can be explained because RVL-CDIP contains a broader panel of document types and is not specialized like BDC. Even though BDC does not contain expense receipts, its global structure is similar to invoices.

Next, we can compare the pre-training tasks. Introducing Numeric Ordering (NO) and Layout Inclusion (LI) tasks also improves the performance over the previously pre-trained model. We observe a 0.32 F1 point improvement on SROIE but only 0.07 on BDC-PO. We suspect the small improvement introduced by the new tasks can be explained because most useful skills to process purchase orders were learned by pre-training on such documents. The new pre-training tasks help more for generalizing on new types of documents.

We also can look at the results on a field per-field basis. We observe that using BDC over RVL-CDIP improved the recognition of all fields except for the dates in BDC-PO. If introducing new training tasks did not improve all fields, we notice that some fields were greatly enhanced like the total amount in SROIE (1.41 F1 points difference). We expected to observe a greater improvement in the total field with the new pre-training tasks. But it does not seem to improve performance much on BDC-PO's total.

Finally it is interesting to compare on Table 3 our results with the published scores of LayoutLM and LayoutLMv2. Our pre-trained model with NO and LI tasks performs better than LayoutLM large which contains 3 times more parameters. However, LayoutLMv2 – which uses both textual and visual information – performance level is still unreachable for a textual-only model.

7 Conclusion

In this work, we showed significant improvements are accessible without introducing more trainable parameters and computational complexity. Only using the base transformer architecture, we achieved a performance that is comparable to the large version which contains 3 times more parameters. Pre-trained models can be further specialized through in-domain datasets and specific pretext training tasks. We demonstrated that by introducing a new collection of business documents and training tasks focusing on documents' layout and number understanding. We showed that performance improvements can be imputed to both pre-training tasks (Numeric Ordering and Layout Inclusion) and new pre-training dataset.

In the future, we will investigate on IE as a Question Answering problem. It has already been proposed in the past [4] as an alternative to Sequence Labeling when fine-tuning models. It should improve the model's generalization capabilities and enable few-shot learning. But nowadays all models are pre-trained, and we would like to study the impact on generalization of a QA-only pre-training.

Bibliography

[1] Denk, T.I., Reisswig, C.: BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding. arXiv:1909.04948 [cs] (Sep 2019), http://arxiv.org/abs/1909.04948, arXiv: 1909.04948

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (May 2019), http://arxiv.org/abs/1810.04805, arXiv: 1810.04805

[3] Gal, R., Ardazi, S., Shilkrot, R.: Cardinal Graph Convolution Framework for Document Information Extraction. In: Proceedings of the ACM Symposium on Document Engineering 2020. pp. 1–11. ACM, Virtual Event CA USA (Sep 2020). https://doi.org/10.1145/3395027.3419584, https://dl.acm.org/doi/10.1145/3395027.3419584

[4] Gardner, M., Berant, J., Hajishirzi, H., Talmor, A., Min, S.: Question Answering is a Format; When is it Useful? arXiv:1909.11291 [cs] (Sep 2019), http://arxiv.org/abs/1909.11291, arXiv: 1909.11291

[5] Garncarek, , Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Graliński, F.: LAMBERT: Layout-Aware language Modeling using BERT for information extraction. arXiv:2002.08087 [cs] (Mar 2020), http://arxiv.org/abs/2002.08087, arXiv: 2002.08087

[6] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520 (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00244, iSSN: 2379-2140

[7] Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., Faddoul, J.B.: Chargrid: Towards Understanding 2D Documents. arXiv:1809.08799 [cs] (Sep 2018), http://arxiv.org/abs/1809.08799, arXiv: 1809.08799

[8] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. arXiv:1603.01360 [cs] (Apr 2016), http://arxiv.org/abs/1603.01360, arXiv: 1603.01360

[9] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 665–666 (2006)

[10] Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., Jagadish, H.V.: Regular Expression Learning for Information Extraction. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. pp. 21–30. Association for Computational Linguistics, Honolulu, Hawaii (Oct 2008), https://www.aclweb.org/anthology/D08-1003

[11] Lin, W., Gao, Q., Sun, L., Zhong, Z., Hu, K., Ren, Q., Huo, Q.: ViBERTgrid: A Jointly Trained Multi-Modal 2D Document Representation for Key Information Extraction from Documents p. 16 (2021)

[12] Liu, X., Gao, F., Zhang, Q., Zhao, H.: Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. arXiv:1903.11279 [cs] (Mar 2019), http://arxiv.org/abs/1903.11279, arXiv: 1903.11279

[13] Lohani, D., Belaïd, A., Belaïd, Y.: An Invoice Reading System Using a Graph Convolutional Network. In: Carneiro, G., You, S. (eds.) Computer Vision – ACCV 2018 Workshops, vol. 11367, pp. 144–158. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-21074-8$_1$2, http://link.springer.com/10.1007/978-3-030-21074-8_12, series Title: Lecture Notes in Computer Science

[14] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs] (Sep 2013), http://arxiv.org/abs/1301.3781, arXiv: 1301.3781

[15] Palm, R.B.: End-to-end information extraction from business documents p. 99 (2019)

[16] Palm, R.B., Winther, O., Laws, F.: CloudScan - A configuration-free invoice analysis system using recurrent neural networks. arXiv:1708.07403 [cs] (Aug 2017), http://arxiv.org/abs/1708.07403, arXiv: 1708.07403

[17] Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1162, http://aclweb.org/anthology/D14-1162

[18] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv:1802.05365 [cs] (Mar 2018), http://arxiv.org/abs/1802.05365, arXiv: 1802.05365

[19] Powalski, R., Borchmann, L., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Palka, G.: Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer p. 18

[20] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training p. 12 (2018)

[21] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs, stat] (Jul 2020), http://arxiv.org/abs/1910.10683, arXiv: 1910.10683

[22] Sage, C., Aussem, A., Eglin, V., Elghazel, H., Espinas, J.: End-to-End Extraction of Structured Information from Business Documents with Pointer-Generator Networks. In: EMNLP 2020 Workshop on Structured Prediction for NLP. Punta Cana (online), Dominican Republic (Nov 2020), https://hal.archives-ouvertes.fr/hal-02958913

[23] Sage, C., Aussem, A., Elghazel, H., Eglin, V., Espinas, J.: Recurrent Neural Network Approach for Table Field Extraction in Business Documents.

In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1308–1313. IEEE (2019)

[24] Sage, C., Douzon, T., Aussem, A., Eglin, V., Elghazel, H., Duffner, S., Garcia, C., Espinas, J.: Data-Efficient Information Extraction from Documents with Pre-trained Language Models. In: Barney Smith, E.H., Pal, U. (eds.) Document Analysis and Recognition – ICDAR 2021 Workshops. pp. 455–469. Springer International Publishing, Cham (2021)

[25] Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields. arXiv:1011.4088 [stat] (Nov 2010), http://arxiv.org/abs/1011.4088, arXiv: 1011.4088

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762 [cs] (Jun 2017), http://arxiv.org/abs/1706.03762, arXiv: 1706.03762

[27] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv:1905.00537 [cs] (Feb 2020), http://arxiv.org/abs/1905.00537, arXiv: 1905.00537

[28] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461 [cs] (Feb 2019), http://arxiv.org/abs/1804.07461, arXiv: 1804.07461

[29] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs] (Jul 2020), http://arxiv.org/abs/1910.03771, arXiv: 1910.03771

[30] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. arXiv:2012.14740 [cs] (May 2021), http://arxiv.org/abs/2012.14740, arXiv: 2012.14740

[31] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining pp. 1192–1200 (Aug 2020). https://doi.org/10.1145/3394486.3403172, http://arxiv.org/abs/1912.13318, arXiv: 1912.13318

[32] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs] (Jan 2020), http://arxiv.org/abs/1906.08237, arXiv: 1906.08237

[33] Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. arXiv:2004.07464 [cs] (Apr 2020), http://arxiv.org/abs/2004.07464, arXiv: 2004.07464