

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

Thibault Douzon, Stefan Duffner, Christophe Garcia and Jérémy Espinas

May 23, 2022

DAS 2022 – Oral Session

Business Documents

REFITECH FRIDGE SERVICES PTE. LTD.
(GST No.: 08883388882)

PURCHASE ORDER

No: 317030
Date: 9/3/2018

PASTOR SINGAPORE PTE. LTD.
110, Pitti Chn Bas Road
Singapore 619792
Tel: 63433333
Fax: 636573151
Attention: Mr. Sabara

Currency: SGD
Delivery: ASAP
Term: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TFP.L.DN15 S.015.F.LK	\$ 101.31	\$ 1013.10
2	4	TFP.L.DN25 S.025.F.LK	\$ 153.04	\$ 612.16
3	2	TFP.L.DN82 S.025.F.LK	\$ 172.97	\$ 345.94
4	10	TFP.L.DN15 S.015.F.LK	\$ 96.55	\$ 965.90
5	4	TFP.L.DN25 S.025.F.LK	\$ 142.07	\$ 571.48
6	2	TFP.L.DN82 S.025.F.LK	\$ 157.48	\$ 314.96
7	3	TFP.E.DN15 S.015.V.E	\$ 210.00	\$ 630.40

Delivery Address:
REFITECH Fridge Services Pte. Ltd.
88A 111, Ubi Yarn Road #
#1-11, Ubi Yarn Plaza 1,
Singapore 489863
Tel: 67430000, Fax: 67430001

Subtotal:	\$ 4,465.99
GST %:	\$ 311.89
Grand Total:	\$ 4,777.88

Remarks: _____

Jeffrey Bob
Authorized Signatory

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

http://www.refitech.com

(a) Purchase order

tan chay yee

*** COPY ***
OJC MARKETING SDN BHD
 ROC NO: 538358-H
 NO 2 & 4, JALAN BAYU 4,
 BANDAR SERI ALAM,
 B1750 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 Person : FATIN
 Bill To : THE PEAK QUARRY WORKS
 Address : ..

Description	Qty	Price	Amount
000000111	1	193.00	193.00 SR

KINGS SAFETY SHOES KWD BOS

Qty: 1	Total Exclude GST:	193.00
	Total GST @5%:	0.00
	Total Inclusive GST:	193.00
	Round Amt:	0.00
	TOTAL:	193.00

VISA CARD	193.00
XXXXXXXXXXXXX4318	
Approval Code:000	

(193.00)

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

(b) Receipt

Figure 1: Document samples from private and public [3] datasets

Information Extraction

REFITECH FROZEN SERVICES PTE. LTD.
Reg. No.: 2008003860

PURCHASE ORDER

No: 017030 Date: 9/3/2018

Currency: SGD Delivery: ASAP Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP L_DN15 S.015.F.L.K	\$ 101.31	\$ 1013.10
2	4	TSP L_DN25 S.025.F.L.K	\$ 153.04	\$ 612.16
3	2	TSP L_DN30 S.030.F.L.K	\$ 172.87	\$ 345.74
4	10	TBF L_DN15 S.015.F.L.K	\$ 98.55	\$ 985.50
5	4	TBF L_DN25 S.025.F.L.K	\$ 142.87	\$ 571.48
6	2	TBF L_DN30 S.030.F.L.K	\$ 157.48	\$ 314.96
7	3	T3BEV_E_DN15 S.015.V.E	\$ 210.80	\$ 632.40

Delivery address:
REFITECH Frozen Services Pte. Ltd.
Bk 311, Ubi Yam Road 5,
1-11, Ubi Yamplex 1,
Singapore 499903
Tel: 67430000, Fax: 67430001

Remarks:

Jeffrey Boh
Authorized Signature

Please send us your order acknowledgement upon receipt of this document.

www.refitech.com.sg

(a) Purchase order

tan chay yee Company

***** COPY *****

OJC MARKETING SDN BHD
ROC NO: 53835B-H
NO 2 & 4, JALAN BAYU 4,
BANDAR SERI ALAM,
81750 MASAI, JOHOR

Tel:07-388 2218 Fax:07-388 8218
Email: np@ojcgroup.com

TAX INVOICE

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

Description	Qty	Price	Amount
000000111	1	193.00	193.00 5R

KINGS SAFETY SHOES KWD B05

Qty:	Total Exclude GST:	193.00
1	Total GST @5%:	0.00
	Total Inclusive GST:	193.00
	Round Amt:	0.00
	TOTAL:	193.00

VISA CARD 193.00
xxxxxx318 Approval Code:000
193.00

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

(b) Receipt

Figure 1: The aim is to extract specific information for each document type

Architectures for Information Extraction

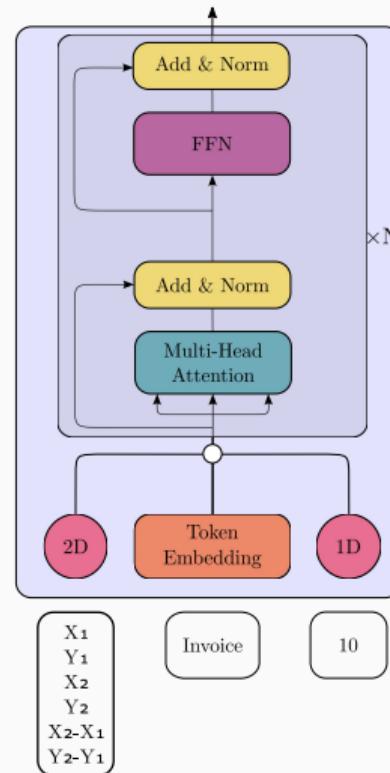


Figure 2: Transformer [4] architecture used for LayoutLM [5, 6] model

Pre-Training Language Models

REFITECH FRIDGE SERVICES PTE. LTD.
GST Reg No: 2008023586G

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 019702
Tel: 688878100
Fax: 688879101
Attn: Mr. Sathas

Date: 9/3/2018
Currency: SGD
Delivery: ASAP
Terms: COD

Item Quantity	Description	Unit Price	Total Price
1	10 TSF-L-DN15 \$0.015/L-K	\$ 101.51	\$ 1013.10
2	4 TSF-L-DN20 \$0.025/L-K	\$ 153.04	\$ 612.16
3	2 TSF-L-DN32 \$0.032/L-K	\$ 172.97	\$ 345.94
4	10 TEF-L-DN15 \$0.015/L-K	\$ 98.55	\$ 985.50
5	4 TEF-L-DN20 \$0.025/L-K	\$ 142.87	\$ 571.48
6	2 TEF-L-DN32 \$0.032/L-K	\$ 157.40	\$ 314.80
7	3 TSBV-F-DH15 \$0.015/V-E	\$ 210.60	\$ 632.40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yam Road 5,
#01-11, Ubi Yamplex 1,
Singapore 489683
Tel: 67430000, Fax: 67430001

Subtotal: \$ 4,455.50
GST 7%: \$ 311.89
Grand Total: \$ 4,767.39

Jeffrey Boh
Authorised Signature

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

www.reftech.com.sg

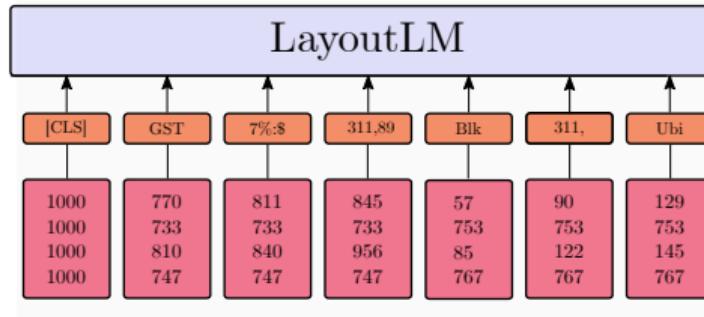


Figure 3: Masked Language Modeling (MLM) [1] diagram. Only part of the sequence is represented.

Pre-Training Language Models

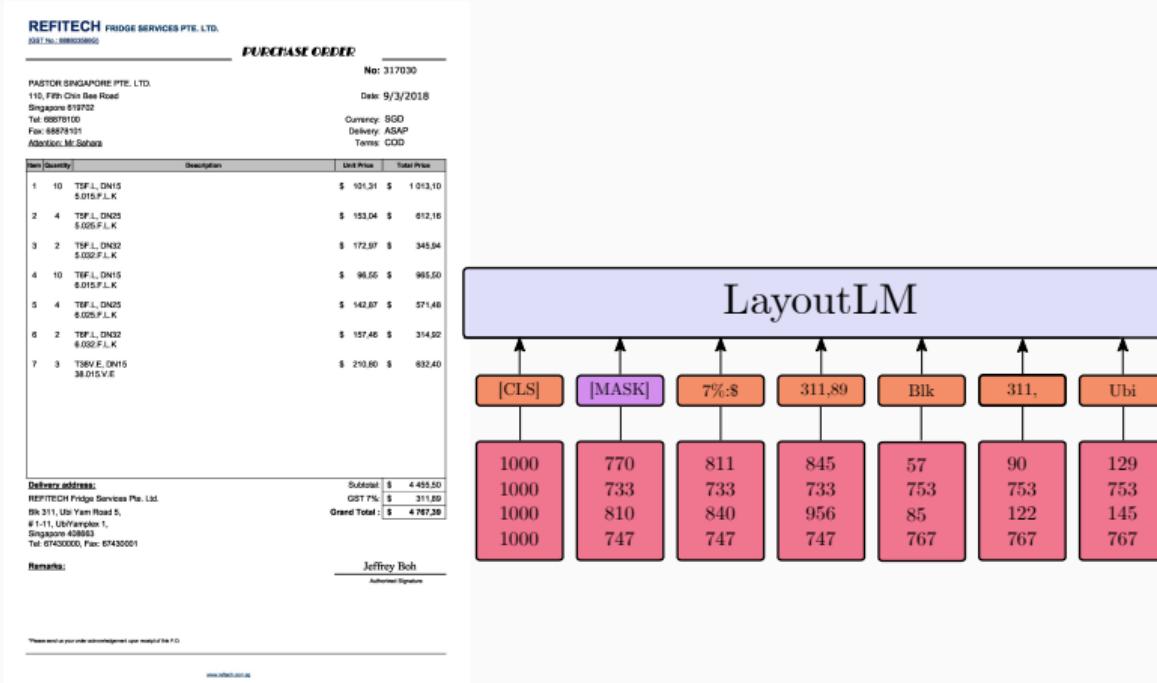


Figure 3: Some tokens are randomly replaced by a special token.

Pre-Training Language Models

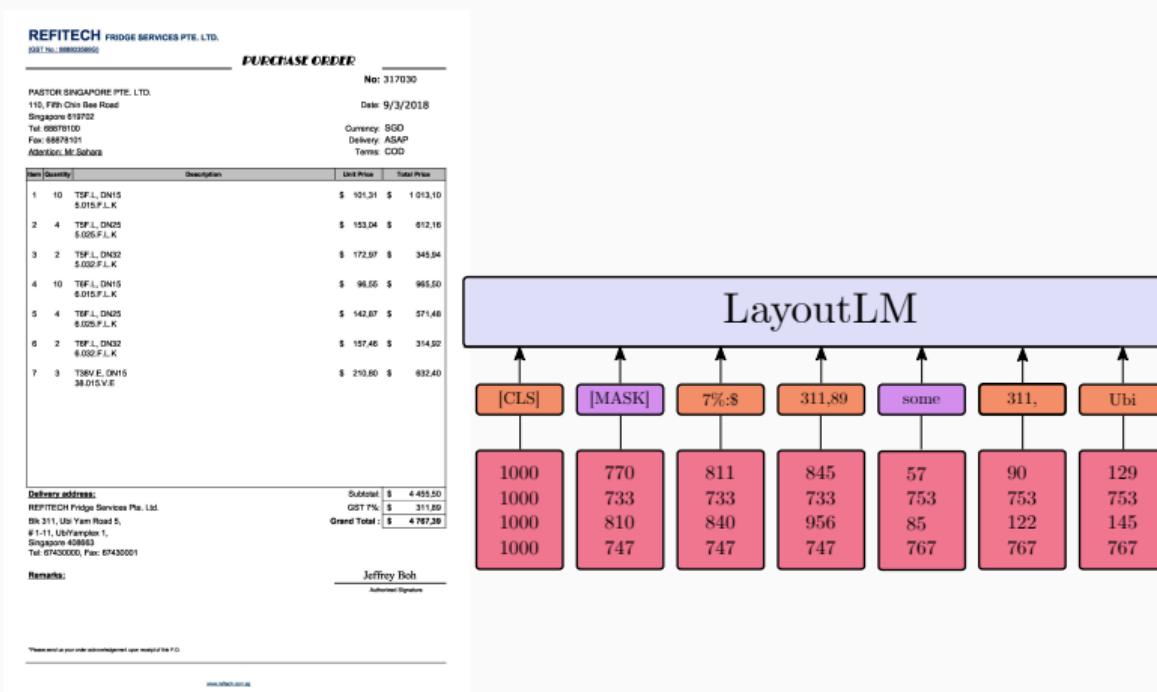


Figure 3: Some other tokens are randomly replaced by other random tokens.

Pre-Training Language Models

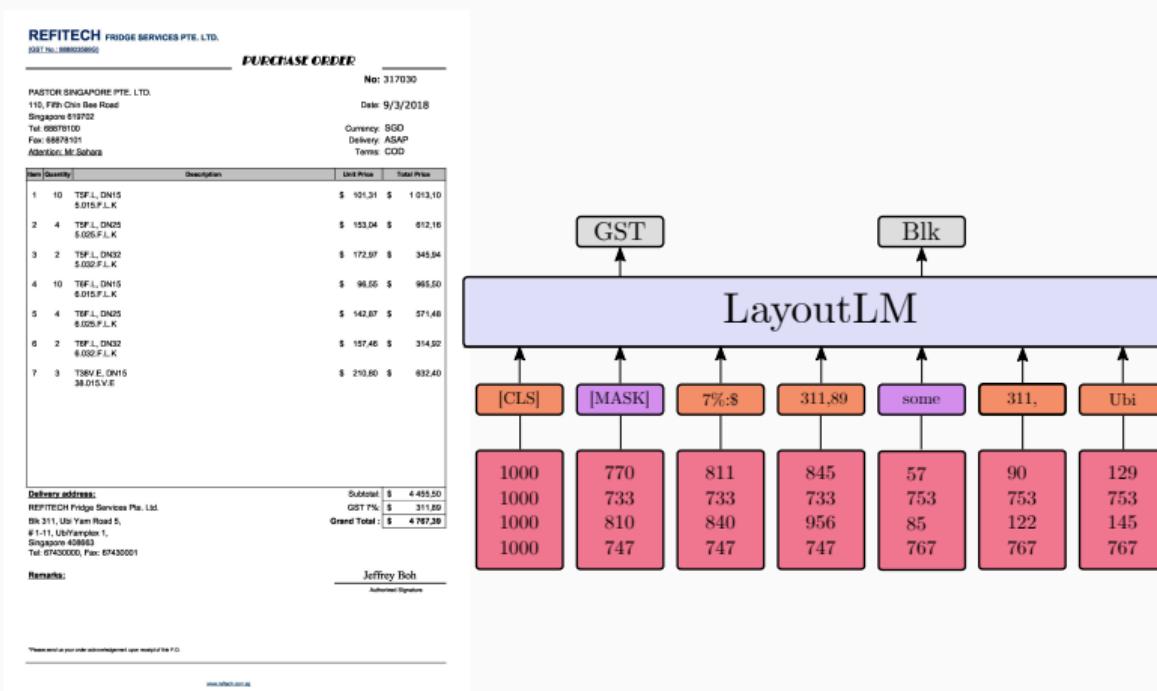


Figure 3: Finally, model is pre-trained by the cross-entropy between prediction logits and initial document's tokens.

Pre-Training Language Models

REFITECH FRIDGE SERVICES PTE. LTD.		PURCHASE ORDER		
(GST No.: 888803568G)				
		No: 317030		
PASTOR SINGAPORE PTE. LTD. 110, Fifth Chin Bee Road Singapore 619702 Tel: 68878100 Fax: 68878101 <u>Attention: Mr Sahara</u>		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

Figure 4: Zoom on a purchase order

Numeric Ordering

REFITECH FROZEN SERVICES PTE. LTD.
 (2017) Reg. No. 2008030002

PURCHASE ORDER

Net: 317030

Date: 9/3/2018

Currency: SGD
 Delivery: ASAP
 Terms: COD

PASTOR SINGAPORE PTE. LTD.
 191, #06-01 Chin Kee Road
 Singapore 018102
 Tel: 65876100
 Fax: 65876101
 Attention: Mr. Sekhar

Item	Quantity	Description	Unit Price	Total Price
1	10	TBP L, DMS5 \$515/L.K	\$ 101.31	\$ 1013.10
2	4	TBP L, DMS5 \$520/L.K	\$ 185.94	\$ 612.16
3	2	TBP L, DMS2 \$532/L.K	\$ 172.97	\$ 345.94
4	10	TBP L, DMS5 \$535/L.K	\$ 98.55	\$ 985.50
5	4	TBP L, DMS5 \$535/L.K	\$ 142.87	\$ 571.48
6	2	TBP L, DMS2 \$535/L.K	\$ 157.49	\$ 314.98
7	3	TBWA, DMS5 \$8312.VSE	\$ 219.80	\$ 659.40

Billing Address:
 REFITECH Frozen Services Pte. Ltd.
 89A, #01-11, Ubi Yarn Road A,
 #01-11, Ubi Yarn Park 1,
 Singapore 408963
 Tel: 65435000, Fax: 65435001

Remarks: _____
 Jeffrey Bob
 Authorized Signatory

*Please send us your acknowledgement upon receipt of this PO.

www.refitech.com

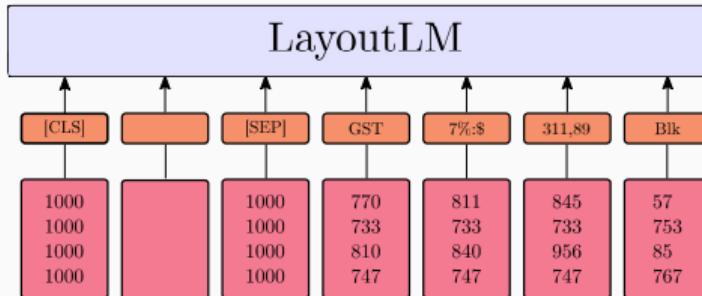


Figure 5: Numeric Ordering (NO) focuses on figures' relative magnitude. Only part of the sequence is represented.

Numeric Ordering

REFITECH FROG SERVICES PTE. LTD.		PURCHASE ORDER	
(Reg. No. 19980000005)		No. 317090	
PASTOR SINGAPORE PTE. LTD.		Date 9/3/2018	
101, Pitt Street West Road			
Singapore 0171702			
Tel: 98871930		Currency SGD	
Fax: 98870101		Delivery ASAP	
Attention: Mr. Saham		Term COD	
Part Number	Description	Unit Price	Total Price
1 10 TFP-L-DM15 S#015-FLK		\$ 161.31	\$ 1613.10
2 4 TFP-L-DM08 S#008-FLK		\$ 183.04	\$ 721.60
3 2 TFP-L-DM03 S#003-FLK		\$ 176.97	\$ 353.94
4 10 TFP-L-DM15 S#015-FLK		\$ 96.55	\$ 965.50
5 4 TFP-L-DM05 S#005-FLK		\$ 142.57	\$ 570.28
6 2 TFP-L-DM02 S#002-FLK		\$ 185.46	\$ 370.92
7 3 TMV-E-DMS 30.015-V-E		\$ 216.80	\$ 652.40
Delivery Address: Refitech Frog Services Pte. Ltd. 80-311, Ubi Parc Road 5, #01-11, Ubi Crescent 1, Singapore 408963 Tel: 67436000, Fax: 67436001		Subtotal	\$ 4495.50
		CST 7%	\$ 311.66
		Grand Total	\$ 4797.16
Remarks:		Jeffrey Bob Authenticated Signature	

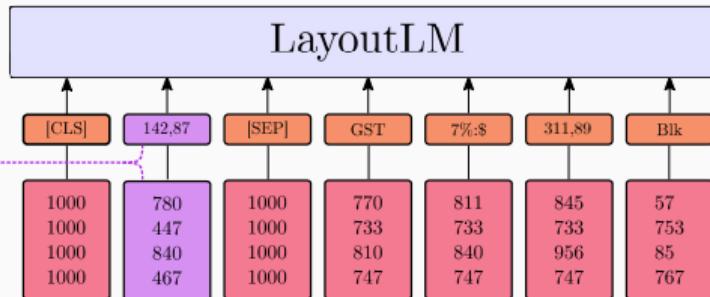


Figure 5: A figure in the document is randomly selected.

Numeric Ordering

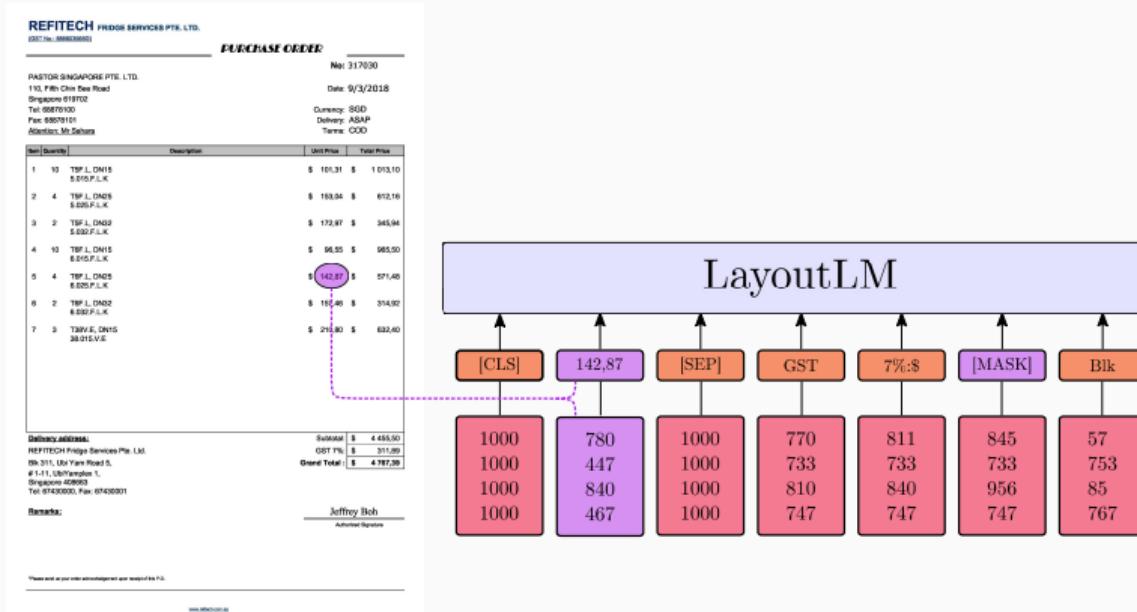


Figure 5: Some tokens and / or positions are randomly masked or replaced.

Numeric Ordering

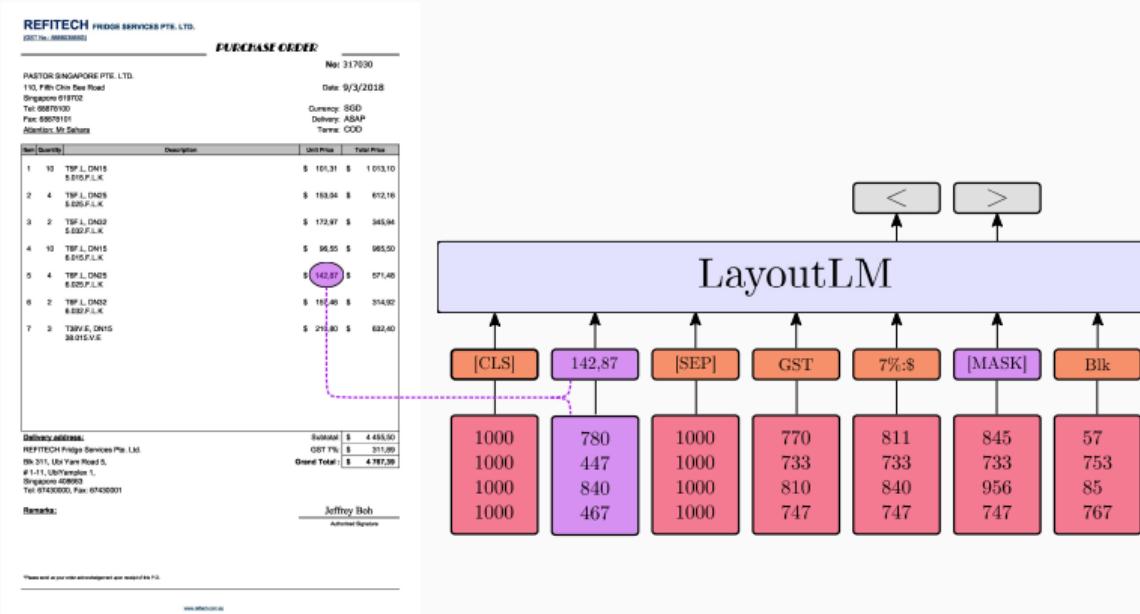


Figure 5: Model is pre-trained on the prediction of relative order between each number and the chosen one.

Layout Inclusion

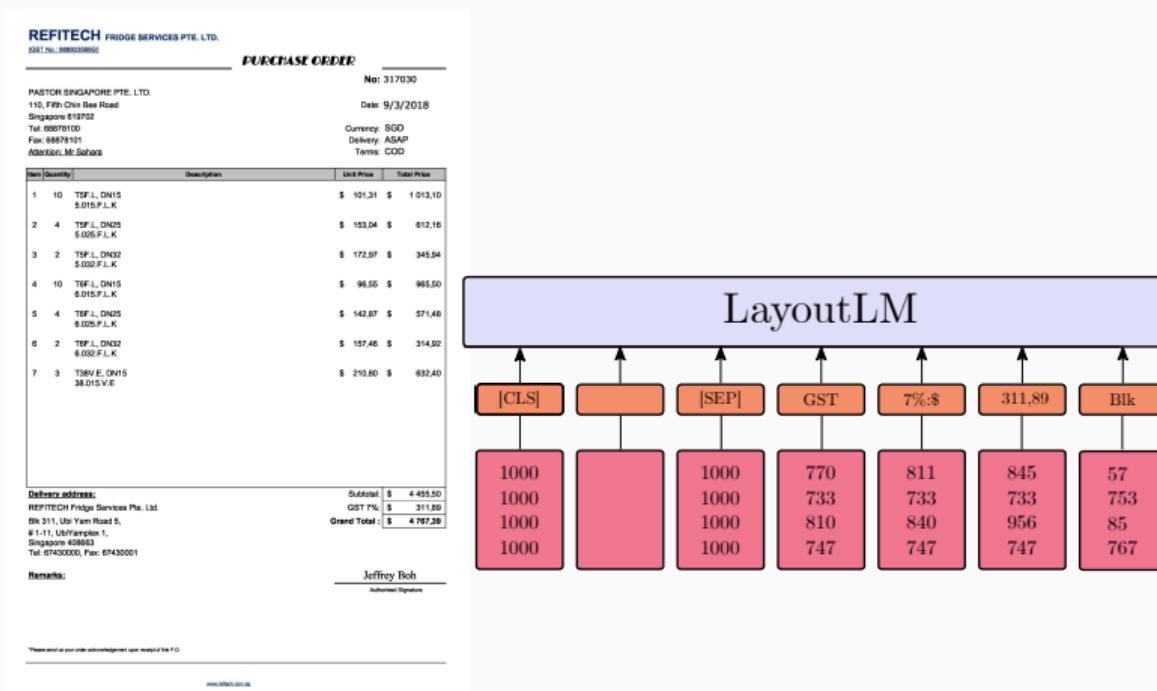


Figure 6: Layout Inclusion (LI) teaches the relative position of tokens. Only part of the sequence is represented.

Layout Inclusion

REFITECH FRIDGE SERVICES PTE. LTD.
GST Reg No: 200802350002

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 019702
Tel: 688878100
Fax: 688879101
Attn: Mr. Sathas

Date: 9/3/2018
Currency: SGD
Delivery: ASAP
Terms: COD

Item Quantity	Description	Unit Price	Total Price
1	10 TSF-L-DN15 \$0.05/F-L,K	\$ 101.31	\$ 1013.10
2	4 TSF-L-DN2 \$0.05/F-L,K	\$ 153.04	\$ 612.16
3	2 TSF-L-DN32 \$0.02/F-L,K	\$ 172.97	\$ 345.94
4	10 TEF-L-DN15 \$0.05/F-L,K	\$ 96.55	\$ 965.50
5	4 TEF-L-DN25 \$0.05/F-L,K	\$ 142.87	\$ 571.48
6	2 TEF-L-DN32 \$0.02/F-L,K	\$ 157.46	\$ 314.92
7	3 TSBV-F-DH15 \$0.015/V-E	\$ 210.80	\$ 632.40
Delivery address: REFITECH Fridge Services Pte. Ltd. Blk 311, Ubi Yarn Plaza 5, #01-11, Ubi Yarn Plaza 5, Singapore 489683 Tel: 67430000, Fax: 67430001		Subtotal \$ 4,455.50 GST 7% \$ 311.89 Grand Total \$ 4,767.39	

Jeffrey Bob
Authorised Signature

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

www.refitech.com.sg

```

graph TD
    A([CLS]) --> B([LAYOUT])
    B --> C([SEP])
    C --> D[GST]
    D --> E[7%:$]
    E --> F[311.89]
    F --> G[Blk]
    B -. dashed line .-> H[1000]
    B -. dashed line .-> I[388]
    B -. dashed line .-> J[1000]
    B -. dashed line .-> K[770]
    B -. dashed line .-> L[811]
    B -. dashed line .-> M[845]
    B -. dashed line .-> N[57]
    H --> O[1000]
    H --> P[1000]
    H --> Q[1000]
    H --> R[1000]
    I --> S[223]
    I --> T[920]
    I --> U[790]
    J --> V[1000]
    J --> W[1000]
    J --> X[1000]
    K --> Y[733]
    K --> Z[810]
    K --> AA[747]
    L --> BB[733]
    L --> CC[840]
    L --> DD[747]
    M --> EE[956]
    M --> FF[747]
    M --> GG[767]
    N --> HH[85]
  
```

Figure 6: A rectangle zone is randomly chosen inside the document's boundaries. It is represented by a **[LAYOUT]** token.

Layout Inclusion

REFITECH FRIDGE SERVICES PTE. LTD.
GST Reg No: 20080235005

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 019702
Tel: 688878100
Fax: 688879101
Attn: Mr. Sathas

Date: 9/3/2018
Currency: SGD
Delivery: ASAP
Terms: COD

Item Quantity	Description	Unit Price	Total Price
1	10 TSF-L-DN15 \$0.05/F-L,K	\$ 101.31	\$ 1013.10
2	4 TSF-L-DN2 \$0.05/F-L,K	\$ 153.04	\$ 612.16
3	2 TSF-L-DN32 \$0.02/F-L,K	\$ 172.97	\$ 345.94
4	10 TEF-L-DN15 \$0.05/F-L,K	\$ 96.55	\$ 965.50
5	4 TEF-L-DN25 \$0.05/F-L,K	\$ 142.87	\$ 571.48
6	2 TEF-L-DN32 \$0.02/F-L,K	\$ 157.46	\$ 314.92
7	3 TSBV-F-DH15 \$0.015/V-E	\$ 210.80	\$ 632.40
		Subtotal: \$ 4455.50	
		GST 7%: \$ 311.89	
		Grand Total: \$ 4767.39	

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yarn Road 5,
#01-11, Ubi Yarnplex 1,
Singapore 489683
Tel: 67430000, Fax: 67430001

Remarks:

Jeffrey Boh
Authorised Signature

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

www.refitech.com.sg

Figure 6: Some tokens and / or positions are randomly masked or replaced.

Layout Inclusion

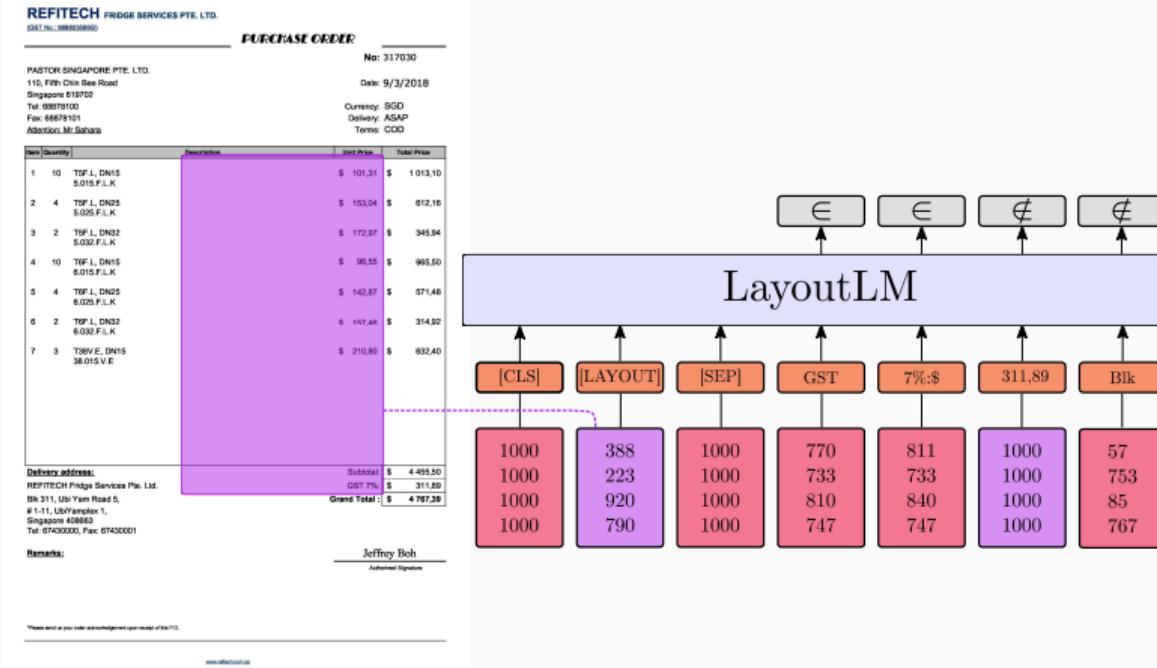


Figure 6: Model predicts for each token if its center is inside or outside the random rectangle.

Pre-Training Datasets

Pre-training datasets

- RVL-CDIP [2], collection of 10M business documents. Pre-trained models available online ;

Pre-Training Datasets

Pre-training datasets

- RVL-CDIP [2], collection of 10M business documents. Pre-trained models available online ;
- Business Document Collection (BDC), 500k invoices and purchase orders from 2018 to today.

Experiments

All experiments evaluated those 3 pre-trained models

- **Masked Language Modeling (MLM)** on **RVL-CDIP**. This model was available online ;

Experiments

All experiments evaluated those 3 pre-trained models

- Masked Language Modeling (MLM) on RVL-CDIP. This model was available online ;
- Masked Language Modeling on the Business Document Collection (BDC) ;

Experiments

All experiments evaluated those 3 pre-trained models

- Masked Language Modeling (MLM) on RVL-CDIP. This model was available online ;
- Masked Language Modeling on the Business Document Collection (BDC) ;
- Masked Language Modeling, Numeric Ordering and Layout Inclusion (MLM+NO+LI) tasks on BDC.

Results

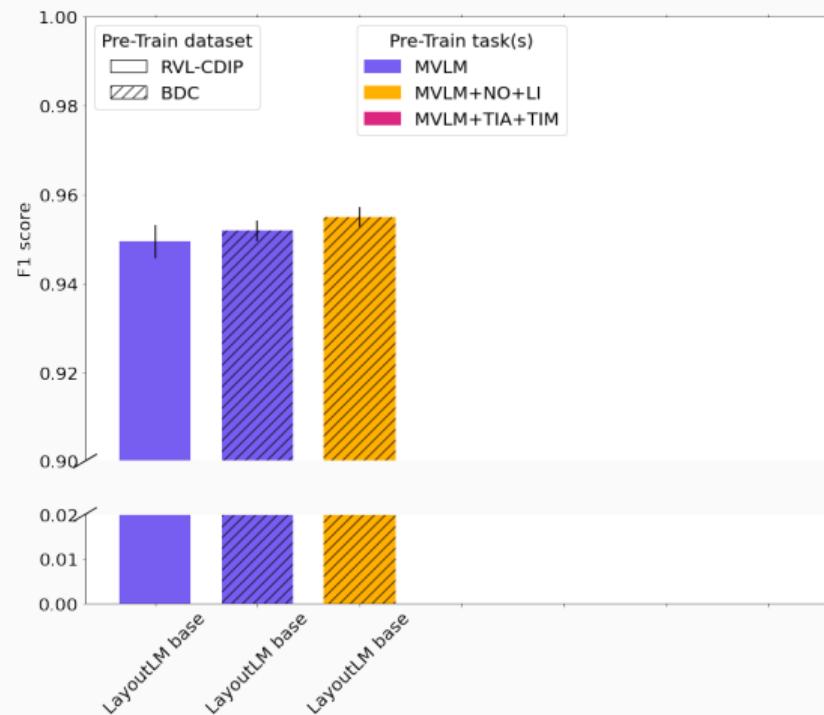


Figure 7: Results on SROIE fine-tuning. Both BDC and the new pre-training tasks improve model performance.

Results

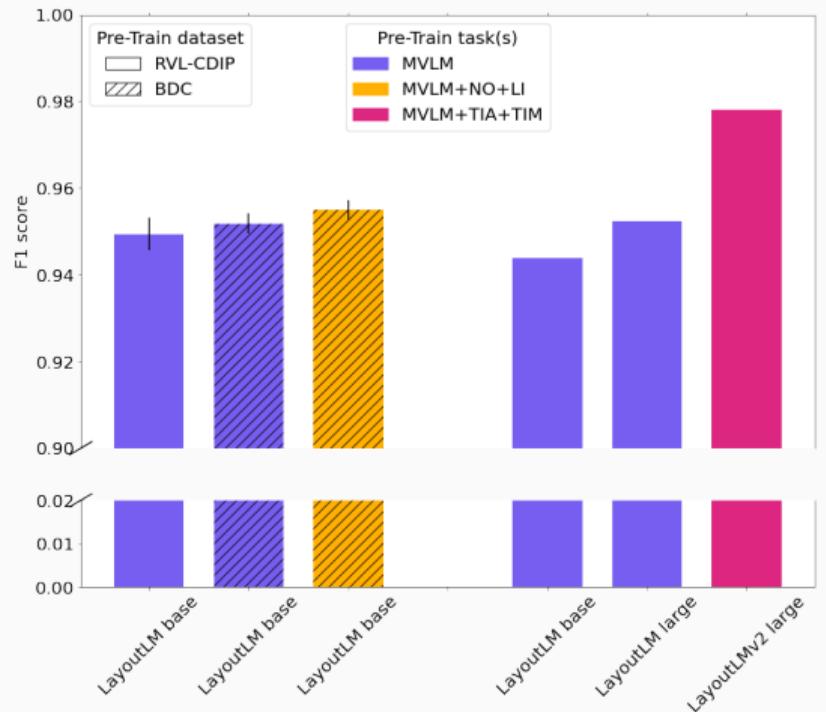


Figure 7: Our pre-trained models (left) outperform original LayoutLM [5, 6] base and large models (right).

Conclusions

- We showed that model's performance on fine-tuning is highly sensitive to pre-training **tasks** and **datasets** ;

Conclusions

- We showed that model's performance on fine-tuning is highly sensitive to pre-training **tasks** and **datasets** ;
- Language model sizes could be reduced – without any performance loss – by elaborating better pre-training adapted to downstream tasks ;

Conclusions

- We showed that model's performance on fine-tuning is highly sensitive to pre-training **tasks** and **datasets** ;
- Language model sizes could be reduced – without any performance loss – by elaborating better pre-training adapted to downstream tasks ;
- More work is needed in order to process very long documents, as current language models are not adapted.

Thanks everyone !

Any questions ?

Paper & contact information

Any in-depth question about this work ? Please contact me !

Thibault Douzon, Stefan Duffner, Christophe Garcia and Jérémie Espinas

thibault.douzon@esker.com

References i

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
arXiv:1810.04805 [cs], May 2019.
arXiv: 1810.04805.
- [2] A. W. Harley, A. Ufkes, and K. G. Derpanis.
Evaluation of deep convolutional nets for document image classification and retrieval.
In *International Conference on Document Analysis and Recognition (ICDAR)*.

References ii

- [3] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar.
ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction.
In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, Sept. 2019.
ISSN: 2379-2140.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin.
Attention Is All You Need.
arXiv:1706.03762 [cs], June 2017.
arXiv: 1706.03762.

References iii

- [5] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou.
LayoutLM: Pre-training of Text and Layout for Document Image Understanding.
Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1192–1200, Aug. 2020.
arXiv: 1912.13318.
- [6] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou.
LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding.
arXiv:2012.14740 [cs], May 2021.
arXiv: 2012.14740.

Presentation theme

Get the source of this theme and the demo presentation from

github.com/matze/mtheme

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

