

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. S88030802)

PURCHASE ORDER

No: 317030

PAETOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619732
Tel: 68870100
Fax: 68870101
Attn: Ms. Sabana

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TRF.L DN15 S.01SF.L.K	\$ 101.31	\$ 1 013.10
2	4	TRF.L DN25 S.02SF.L.K	\$ 153.04	\$ 612.16
3	2	TRF.L DN32 S.02SF.L.K	\$ 172.97	\$ 345.94
4	10	TRF.L DN15 S.01SF.L.K	\$ 96.55	\$ 965.50
5	4	TRF.L DN25 S.02SF.L.K	\$ 142.87	\$ 571.48
6	2	TRF.L DN32 S.02SF.L.K	\$ 157.49	\$ 314.92
7	3	T26V.E. DN15 38.015V.E	\$ 210.80	\$ 632.40

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

Subtotal \$ 4 455.90
GST 7% \$ 311.89
Grand Total \$ 4 767.79

Remarks: _____
Jeffrey Boh
Authorised Signature

*This work is your own responsibility and not that of P&G.
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,
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 Person : FATIN
Bill To : **THE PEAK QUARRY WORKS**
Address : . .

Description	Qty	Price	Amount
000000111	1	193.00	193.00 SR
KINGS SAFETY SHOES KWD B05			
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			

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 FRIDGE SERVICES PTE. LTD.
GST No. 80005963

PURCHASE ORDER

No: **317030** — PO Number
Date: **9/3/2018** — Date

Currency: SGD
Delivery: ASAP
Terms: COD

PASTOR SINGAPORE PTE. LTD.
115, Fifth Cross Street Road
Singapore 051002
Tel: 68678100
Fax: 98878101
Attention: Mr. Sathish

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DN15 S.D15.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DN25 S.D25.F.L.K	\$ 193.04	\$ 812.16
3	2	TSP_L_DN32 S.D32.F.L.K	\$ 172.87	\$ 345.94
4	10	TSP_L_DN15 S.D15.F.L.K	\$ 98.55	\$ 985.50
5	4	TSP_L_DN25 S.D25.F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DN32 S.D32.F.L.K	\$ 157.48	\$ 314.92
7	3	T3BEVE_DN15 38.S15.V.E	\$ 210.80	\$ 632.40

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

Delivery address:
REFITECH Fridge Services Pte. Ltd.
030-311, Ulu-Yan Road 5,
#11-11, UluYan@paleo 1,
Singapore 459093
Tel: 07430300, Fax: 07430301

Remarks: Jeffrey Boh
Authorized Signature

(a) Purchase order

tan chay yee — Company

*** COPY ***

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

Address

TAX INVOICE

Invoice No : PEGIV-1030765
Date : **15/01/2019 11:05:10 AM** — Date
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 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 — Total

VISA CARD 193.00
XXXXXXXXXXXX4318
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

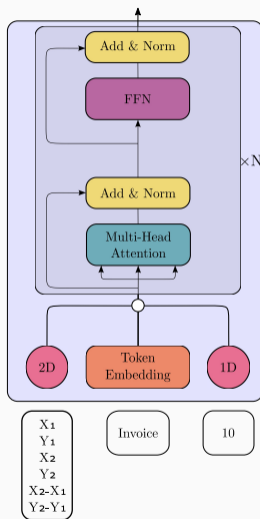


Figure 2: Transformer [4] architecture used for LayoutLM [5, 6] model

Pre-Training Language Models

REFITECH PRIDGE SERVICES PTE. LTD.
GST No. 88828963

PURCHASE ORDER

Ref: 317030

DATE: 9/3/2018

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 051102
Tel: 68876100
Fax: 68879101
Address: Mr. Sohwa

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L DN15 S 015.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSF L DN25 S 025.F.L.K	\$ 153.04	\$ 612.16
3	2	TSF L DN32 S 032.F.L.K	\$ 172.97	\$ 345.94
4	10	TSF L DN15 S 015.F.L.K	\$ 96.55	\$ 965.50
5	4	TSF L DN25 S 025.F.L.K	\$ 142.87	\$ 571.48
6	2	TSF L DN32 S 032.F.L.K	\$ 157.46	\$ 314.92
7	3	TSFV E DN15 38.015.V.E	\$ 210.80	\$ 632.40

Delivery address: REFITECH Pridge Services Pte. Ltd.
816 311, 150 Yew Road 5,
#1-11, Ubi/Prampok 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 read all your order acknowledgment upon receipt of the P.O.

www.refitech.com.sg

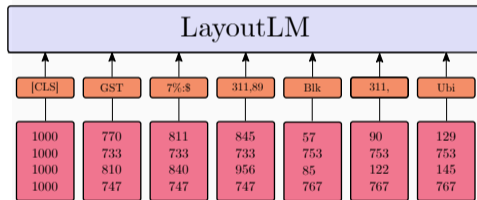


Figure 3: Masked Language Modeling (MLM) [1] diagram. Only part of the sequence is represented.

Pre-Training Language Models

REFITECH BRIDGE SERVICES PTE. LTD.
GST No. 88828963

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 051022
Tel: 69876100
Fax: 68879101
Address: Mr. Sohwa

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 S.015.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 S.025.F.L.K	\$ 153.04	\$ 612.16
3	2	TSP_L_DM32 S.032.F.L.K	\$ 172.97	\$ 345.94
4	10	TSP_L_DM15 S.015.F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 S.025.F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM32 S.032.F.L.K	\$ 157.46	\$ 314.92
7	3	TSHV E. DR15 38.015.V.E	\$ 293.80	\$ 881.40

Delivery address: REFITECH Bridge Services Pte. Ltd.
816 311, 150 Yew Road 5,
#1-11, Ubi/Anchor 1,
Singapore 408963
Tel: 67420000, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signature

*Please read our your order acknowledgment upon receipt of the P.O.

www.refitech.com.sg

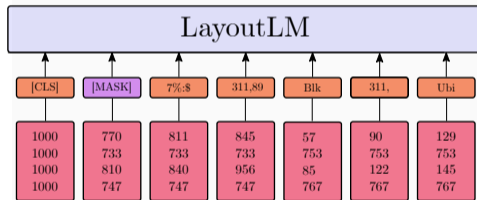


Figure 3: Some tokens are randomly replaced by a special token.

Pre-Training Language Models

REFITECH BRIDGE SERVICES PTE. LTD.
GST No. 88828963

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 051022
Tel: 68876100
Fax: 68879101
Address: Mr. Sohwa

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 \$ 015.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 \$ 025.F.L.K	\$ 153.04	\$ 612.16
3	2	TSP_L_DM32 \$ 032.F.L.K	\$ 172.87	\$ 345.84
4	10	TSP_L_DM15 \$ 015.F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 \$ 025.F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM32 \$ 032.F.L.K	\$ 157.46	\$ 314.92
7	3	TSHV E. DR15 38.015.V.E	\$ 293.80	\$ 881.40

Delivery address: REFITECH Bridge Services Pte. Ltd.
816 311, 150 Yew Road 5,
#1-11, Ubi/Anchor 1,
Singapore 408963
Tel: 67430000, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signature

*Please read our user guide and implement user manual of the P.O.

www.refitech.com

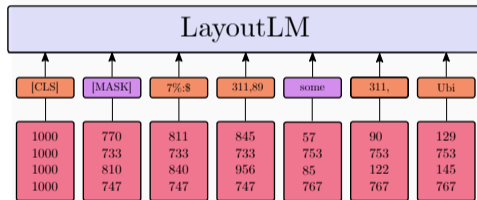


Figure 3: Some other tokens are randomly replaced by other random tokens.

Pre-Training Language Models

REFITECH PRIDGE SERVICES PTE. LTD.
GST No. 88828963

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 051022
Tel: 68876100
Fax: 68879101
Address: Mr. Sohwa

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TFP L DN15 \$ 015 P.L.K	\$ 101.31	\$ 1 013.10
2	4	TFP L DN25 \$ 025 P.L.K	\$ 153.04	\$ 612.16
3	2	TFP L DN32 \$ 032 P.L.K	\$ 172.97	\$ 345.94
4	10	TFP L DN15 \$ 015 P.L.K	\$ 96.55	\$ 965.50
5	4	TFP L DN25 \$ 025 P.L.K	\$ 142.87	\$ 571.48
6	2	TFP L DN32 \$ 032 P.L.K	\$ 157.46	\$ 314.92
7	3	TSHV E DN15 38.015 V.E	\$ 219.80	\$ 659.40

Delivery address: REFITECH Pridge Services Pte. Ltd.
816 311, 150 Yern Road 5,
#1-11, Ubi/Anchor 1,
Singapore 408963
Tel: 67430000, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signature

*Please read our user guide and development user manual at the P.O.

www.refitech.com.sg

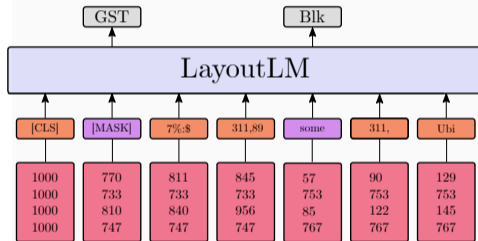


Figure 3: Finally, model is pre-trained by the cross-entropy between prediction logits and initial document's tokens.

Pre-Training Language Models

REFITECH FRIDGE SERVICES PTE. LTD. (GST No.: 888803566G)		PURCHASE ORDER		
<hr/>		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 FRIDGE SERVICES PTE. LTD.
2377161-88888888

PURCHASE ORDER

Ref: 317030

PASTOR SINGAPORE PTE. LTD. Date: 9/3/2018
110, Fifth Cross Street
Singapore 018702
Tel: 65879100 Currency: SGD
Fax: 65879101 Delivery: ASAP
Attention: Mr. Sathya Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 \$ 103.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 \$ 103.F.L.K	\$ 188.04	\$ 752.16
3	2	TSP_L_DM20 \$ 103.F.L.K	\$ 173.97	\$ 347.94
4	10	TSP_L_DM15 \$ 103.F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 \$ 103.F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM20 \$ 103.F.L.K	\$ 187.40	\$ 374.80
7	3	TSPV_E_DM15 38.01515/G	\$ 212.80	\$ 638.40

Delivery address: Ref: Ref: \$ 4 455.50
REFITECH Fridge Services Pte. Ltd. GST 1% \$ 311.20
39, 311, Ulu Yank Road S. Grand Total: \$ 4 767.28
1-11, Ulu Yam Road 1,
Singapore 670802
Tel: 67430005, Fax: 67430011

Remarks: Jeffrey Boh
Authorized Signatory

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

www.refitech.com

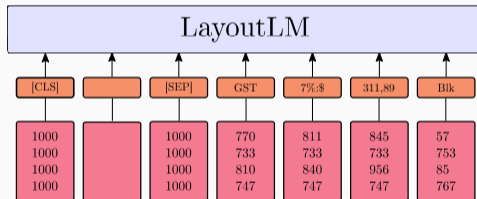


Figure 5: Numeric Ordering (NO) focuses on figures' relative magnitude. Only part of the sequence is represented.

Numeric Ordering

REFITECH FRIDGE SERVICES PTE. LTD.
 (2017040000000)

PURCHASE ORDER

Ref: 317030
 Date: 9/3/2018

PASTOR SINGAPORE PTE. LTD.
 110, Fifth Cross Street
 Singapore 018702
 Tel: 68879100
 Fax: 68879101
 Attention: Mr. Sathya

Currency: SGD
 Delivery: ASAP
 Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 8.025.F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 8.025.F.L.K	\$ 189.04	\$ 762.16
3	2	TSP_L_DM22 8.032.F.L.K	\$ 173.97	\$ 347.94
4	10	TSP_L_DM15 8.010.F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 8.025.F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM22 8.032.F.L.K	\$ 185.49	\$ 370.98
7	3	TSPV_E_DM15 28.012.V.G	\$ 212.80	\$ 638.40

Delivery address:
 REFITECH Fridge Services Pte. Ltd.
 09-211, Ulu Yam Road 5,
 #1-11, Ulu Yam Plaza 1,
 Singapore 670802
 Tel: 67430005, Fax: 67430001

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

Remarks: Jeffrey Boh
 Authorised Signature

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

www.refitech.com.sg

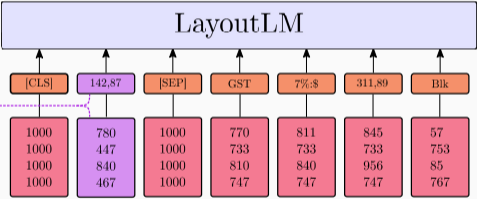


Figure 5: A figure in the document is randomly selected.

Numeric Ordering

REFITECH FRIDGE SERVICES PTE. LTD.
23779499999999

PURCHASE ORDER

Ref: 317030
Date: 9/3/2018

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street Road
Singapore 018702
Tel: 65879100
Fax: 65879101
Attention: Mr. Subram

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 S.035/F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 S.025/F.L.K	\$ 188.04	\$ 752.16
3	2	TSP_L_DM20 S.032/F.L.K	\$ 173.97	\$ 347.94
4	10	TSP_L_DM15 S.035/F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 S.025/F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM20 S.032/F.L.K	\$ 185.49	\$ 370.98
7	3	TSPV_E_DM15 38.01212.0	\$ 212.80	\$ 638.40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
09-311, Ulu Yam Road S.
#1-11, Ulu Yam Road 1,
Singapore 670802
Tel: 67430005, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signatory

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

www.refitech.com

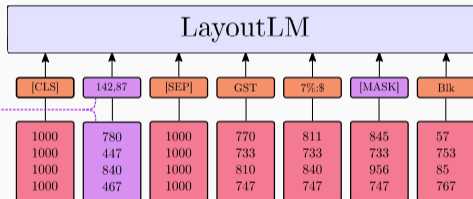


Figure 5: Some tokens and / or positions are randomly masked or replaced.

Numeric Ordering

REFITECH FRIDGE SERVICES PTE. LTD.
237796, 88888888

PURCHASE ORDER

Ref: 317930 Date: 9/3/2018

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 018702
Tel: 68879100
Fax: 68879101
Attention: Mr. Sathya

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP_L_DM15 8.025 F.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP_L_DM25 8.025 F.L.K	\$ 188.04	\$ 752.16
3	2	TSP_L_DM22 8.032 F.L.K	\$ 173.97	\$ 347.94
4	10	TSP_L_DM15 8.015 F.L.K	\$ 96.55	\$ 965.50
5	4	TSP_L_DM25 8.025 F.L.K	\$ 142.87	\$ 571.48
6	2	TSP_L_DM22 8.032 F.L.K	\$ 185.49	\$ 370.98
7	3	TSPV_E_DM15 38.015 V.G	\$ 212.80	\$ 638.40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
09, 211, Ulu Yam Road S.
#1-11, Ulu Yam Road 1,
Singapore 670802
Tel: 67430005, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signatory

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

www.refitech.com

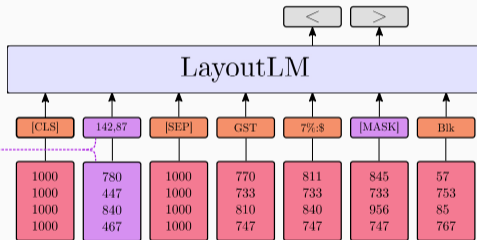


Figure 5: Model is pre-trained on the prediction of relative order between each number and the chosen one.

Layout Inclusion

REFITECH PRIDGE SERVICES PTE. LTD.
GST No. 88828963

PURCHASE ORDER

Ref: 317030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Cross Street
Singapore 051022
Tel: 68876100
Fax: 68879101
Attention: Mr. Sohwa

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L DN15 S 015 P L K	\$ 101.31	\$ 1 013.10
2	4	TSF L DN25 S 025 F L K	\$ 153.04	\$ 612.16
3	2	TSF L DN32 S 032 F L K	\$ 172.97	\$ 345.94
4	10	TSF L DN15 S 015 P L K	\$ 96.55	\$ 965.50
5	4	TSF L DN25 S 025 F L K	\$ 142.87	\$ 571.48
6	2	TSF L DN32 S 032 F L K	\$ 157.46	\$ 314.92
7	3	TSFV E DN15 38.015 V E	\$ 219.80	\$ 659.40

Delivery address: REFITECH Pridge Services Pte. Ltd.
816 311, 150 Yew Road 5,
#1-11, Ubi/Anchor 1,
Singapore 408963
Tel: 67430000, Fax: 67430001

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

Remarks: Jeffrey Boh
Authorized Signature

*Please read all your order acknowledgements upon receipt of the P.O.

www.refitech.com.sg

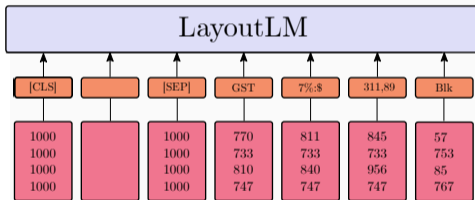


Figure 6: Layout Inclusion (LI) teaches the relative position of tokens. Only part of the sequence is represented.

Layout Inclusion

REFITECH PRIDGE SERVICES PTE. LTD.
 0311 NL 88828963

PURCHASE ORDER

REFITECH PRIDGE SERVICES PTE. LTD.
 110, Fifth Cross Street
 Singapore 051022
 Tel: 68876100
 Fax: 68879101
 Attention: Mr. Sohwa

No: 317030
 Date: 9/3/2018
 Currency: SGD
 Delivery: ASAP
 Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TFP L DN15 S 015 P L K	\$ 101.31	\$ 1 013.10
2	4	TFP L DN25 S 025 F L K	\$ 153.04	\$ 612.16
3	2	TFP L DN32 S 032 F L K	\$ 172.87	\$ 345.84
4	10	TFP L DN15 S 015 P L K	\$ 96.55	\$ 965.50
5	4	TFP L DN25 S 025 F L K	\$ 142.87	\$ 571.48
6	2	TFP L DN32 S 032 F L K	\$ 157.46	\$ 314.92
7	3	TSHV E DR15 38.015 V E	\$ 270.80	\$ 812.40

Delivery address: REFITECH Pridge Services Pte. Ltd.
 816 311, 5th Yam Road 5,
 #1-11, Ubi/Prampok 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 read our user guide at www.refitech.com.sg for details of this P.O.

www.refitech.com.sg

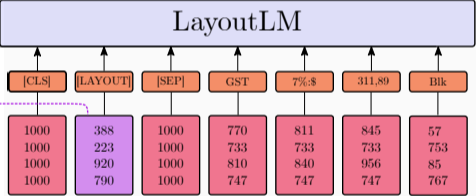


Figure 6: A rectangle zone is randomly chosen inside the document’s boundaries. It is represented by a [LAYOUT] token.

Layout Inclusion

REFITECH PRIDGE SERVICES PTE. LTD.
 0317.NL.88828963

PURCHASE ORDER

Ref: 317030

PASTOR SINGAPORE PTE. LTD.
 110, Fifth Cross Sea Road
 Singapore 018702
 Tel: 68876100
 Fax: 68879101
 Address: Mr. Sohwa

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF.L.DM15 S.015.P.L.K	\$ 101.31	\$ 1 013.10
2	4	TSF.L.DM25 S.025.F.L.K	\$ 153.04	\$ 612.16
3	2	TSF.L.DM32 S.032.F.L.K	\$ 172.87	\$ 345.84
4	10	TSF.L.DM15 S.015.P.L.K	\$ 96.55	\$ 965.50
5	4	TSF.L.DM25 S.025.F.L.K	\$ 142.87	\$ 571.48
6	2	TSF.L.DM32 S.032.F.L.K	\$ 157.46	\$ 314.92
7	3	TSBV.E.DM15 38.015.V.E	\$ 270.80	\$ 812.40

Delivery address: REFITECH Pridge Services Pte. Ltd.
 #11-11, Ubi/Anchor 1,
 Singapore 408863
 Tel: 67430000, Fax: 67430001

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

Remarks: Jeffrey Boh
 Authorized Signature

*Please read all your order acknowledgment upon receipt of the P.O.

www.refitech.com.sg

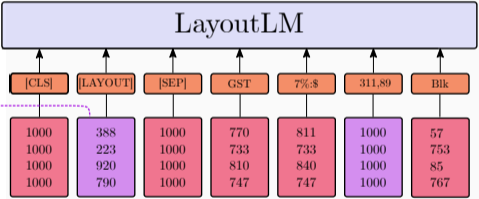


Figure 6: Some tokens and / or positions are randomly masked or replaced.

Layout Inclusion

REFITECH FRIDGE SERVICES PTE. LTD.
 (SST No.: 88826865)

PURCHASE ORDER

No: 317030

PASTOR SINGAPORE PTE. LTD.
 110, Fifth Chin Bee Road
 Singapore 618702
 Tel: 68878100
 Fax: 68878101
 Attention: Mr. Subash

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSP L, DN15 S.015 P.L.K	\$ 101.31	\$ 1 013.10
2	4	TSP L, DN25 S.025 P.L.K	\$ 153.04	\$ 612.16
3	2	TSP L, DN32 S.032 P.L.K	\$ 172.37	\$ 344.74
4	10	TWP L, DN15 S.015 P.L.K	\$ 95.55	\$ 955.50
5	4	TWP L, DN25 S.025 P.L.K	\$ 142.87	\$ 571.48
6	2	TWP L, DN32 S.032 P.L.K	\$ 147.44	\$ 294.88
7	3	TSPV E, DN15 38.015 V.E	\$ 210.80	\$ 632.40

Delivery address: Subash \$ 4 455.50
 REFITECH Fridge Services Pte. Ltd.
 #1-11, Ulu Yam Road 5,
 Singapore 488663
 Tel: 67430000, Fax: 67430001
 GST 7%: \$ 311.89
Grand Total: \$ 4 767.39

Remarks: Jeffrey Boh
 Authorized Signatory

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

www.refitech.com

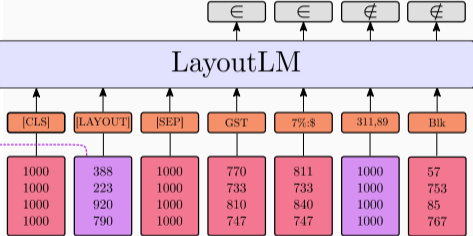


Figure 6: Model predicts for each token if its center is inside or outside the random rectangle.

Pre-training datasets

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

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.

All experiments evaluated those 3 pre-trained models

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

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)** ;

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

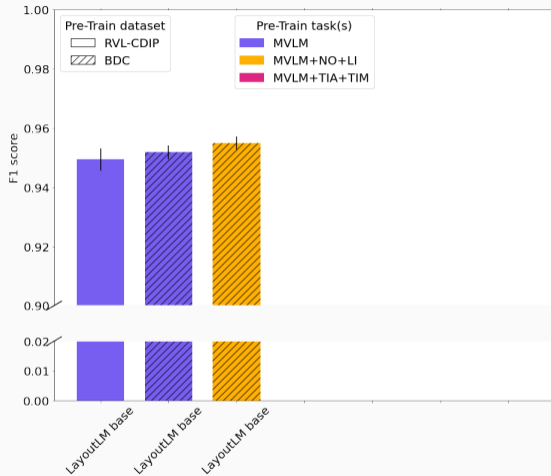


Figure 7: Results on SROIE fine-tuning. Both BDC and the new pre-training tasks improve model performance.

Results

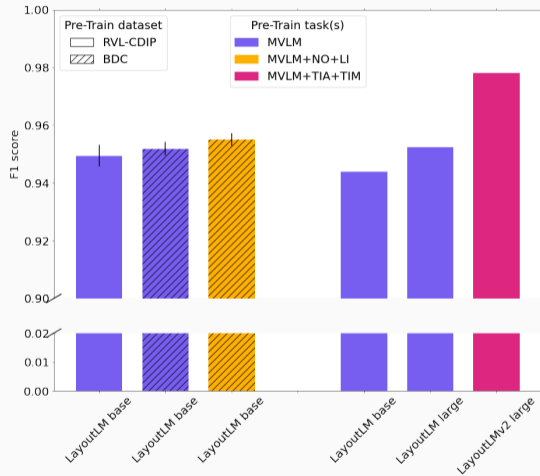


Figure 7: Our pre-trained models (left) outperform original LayoutLM [5, 6] base and large models (right).

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 Jérémy Espinas

thibault.douzon@esker.com

Code and models are available on

github.com/thibaultdouzon/business-document-pre-training.git

- [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).

- [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.

- [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.

