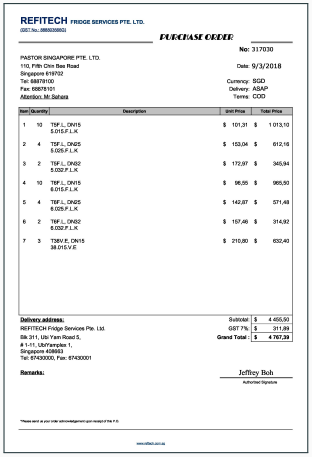
Improving Information Extraction on Business Documents with Specic Pre-Training Tasks

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.: 200302066G)

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

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yam Road 1,
1-11, Ubi Yamplaza 1,
Singapore 408693
Tel: 67430000, 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 this P.O.

www.refitech.com.sg

(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 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	

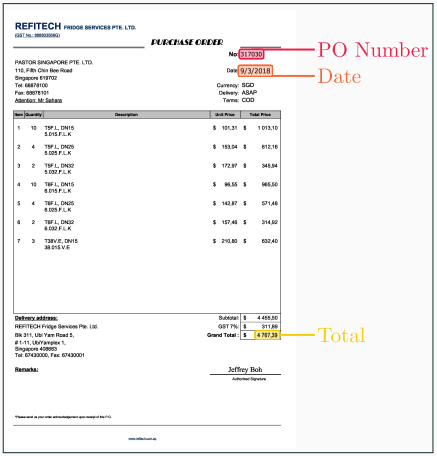
193.00

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

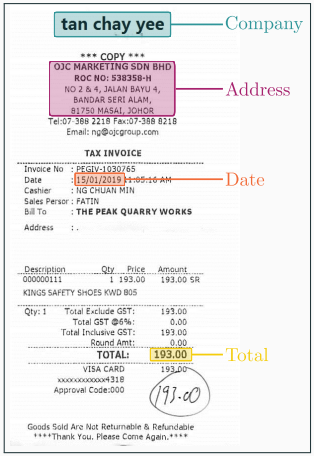
(b) Receipt

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

Information Extraction



(a) Purchase order



(b) Receipt

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

Architectures for Information Extraction

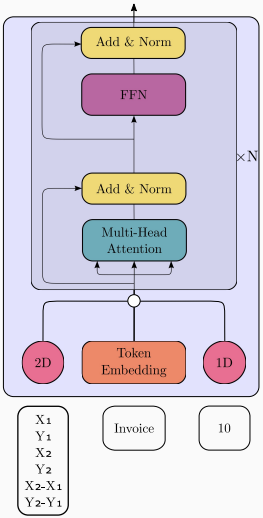
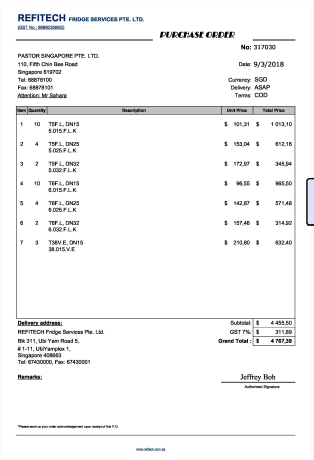


Figure 2: Transformer [4] architecture used for LayoutLM [5, 6] model

Pre-Training Language Models



REFITECH FRIDGE SERVICES PTE. LTD.

GST No.: 888833888G

PURCHASE ORDER

No: 3170130

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68978100
Fax: 68978101
Attention: Mr Sahara

Date: 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L. DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	TSF L. DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	TSF L. DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	TSF L. DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	TSF L. DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	TSF 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, LSV Complex 1,
Singapore 408863
Tel: 67400000, 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

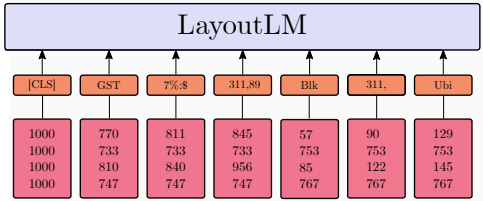
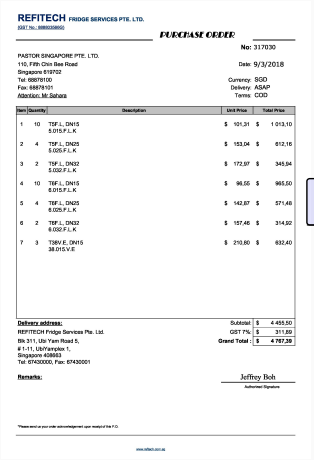


Figure 3: Masked Language Modeling (MLM) [1] diagram. Only part of the sequence is represented.

Pre-Training Language Models



REFITECH FRIDGE SERVICES PTE. LTD.

GST No.: 888833888G

PURCHASE ORDER

No: 3170130

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68978100
Fax: 68878101
Attention: Mr Sahara

Date: 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L, DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	TSF L, DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	TSF L, DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	TSF L, DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	TSF L, DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	TSF 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, LSV Complex 1,
Singapore 408863
Tel: 67400000, 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

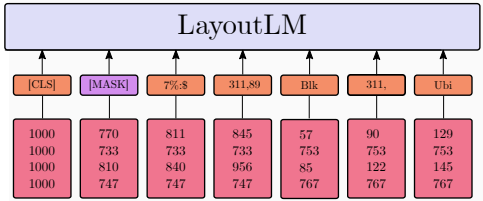


Figure 3: Some tokens are randomly replaced by a special token.

Pre-Training Language Models

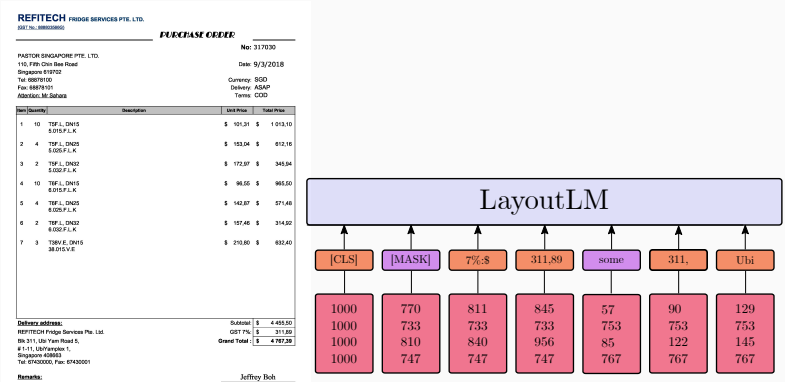
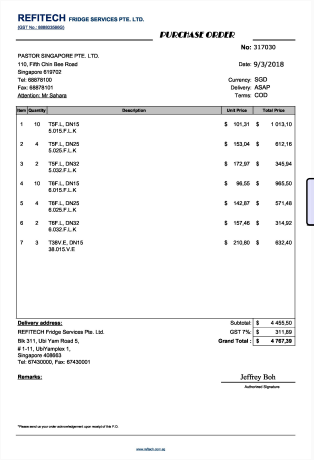


Figure 3: Some other tokens are randomly replaced by other random tokens.

Pre-Training Language Models



REFITECH FRIDGE SERVICES PTE. LTD.

GST No.: 888833888G

PURCHASE ORDER

No: 3170130

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68978100
Fax: 68978101
Attention: Mr Sahara

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L. DN15 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	TSF L. DN25 5.025.F.L.K	$ 153,04	$ 612,16
3	2	TSF L. DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	TSF L. DN15 6.015.F.L.K	$ 96,55	$ 965,50
5	4	TSF L. DN25 6.025.F.L.K	$ 142,87	$ 571,48
6	2	TSF 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, LSV Complex 1,
Singapore 408863
Tel: 67400000, 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

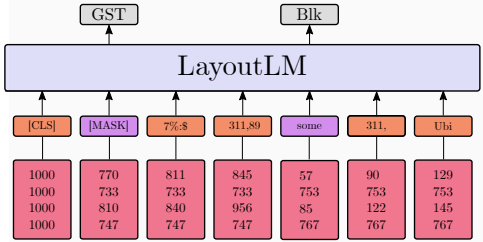
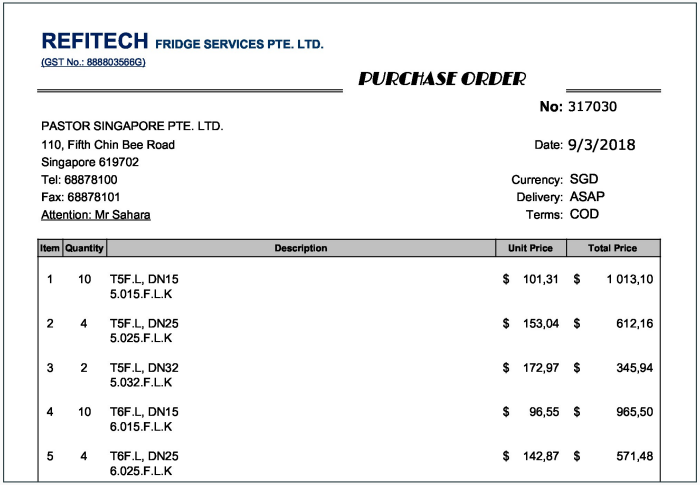


Figure 3: Finally, model is pre-trained by the cross-entropy between prediction logits and intial document's tokens.

Pre-Training Language Models



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

PURCHASE ORDER

No: 317030

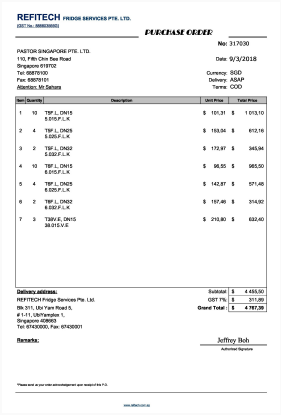
PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68878100
Fax: 68878101
Attention: Mr Sahara

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

Figure 4: Zoom on a purchase order

Numeric Ordering



REFITECH FRIDGE SERVICES PTE. LTD.
(GST No.: 200823456D)

PURCHASE ORDER

No: 317630

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Swe Road
Singapore 618732
Tel: 68876100
Fax: 68876101
Attention: Mr Sahara

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

Item	Quantity	Description	Unit Price	Total Price
1	10	TBF L, DN15 S.015.F.L.K	$ 101,31	$ 1 013,10
2	4	TBF L, DN25 S.025.F.L.K	$ 153,04	$ 612,16
3	2	TBF L, DN32 S.032.F.L.K	$ 172,97	$ 345,94
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, DN32 S.032.F.L.K	$ 197,46	$ 314,92
7	3	TBW-E, DN15 38.015.V.E	$ 210,80	$ 632,40

Delivery address:
REFITECH Fridge Services Pte. Ltd.
Blk 311, Ubi Yan Road 5,
1-11, Ubi Paradise 1,
Singapore 408963
Tel: 67430000, 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 this P.O.

www.refitech.sg

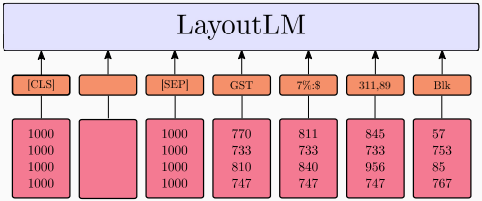


Figure 5: Numeric Ordering (NO) focuses on figures' relative magnitude. Only part of the sequence is represented.

Numeric Ordering

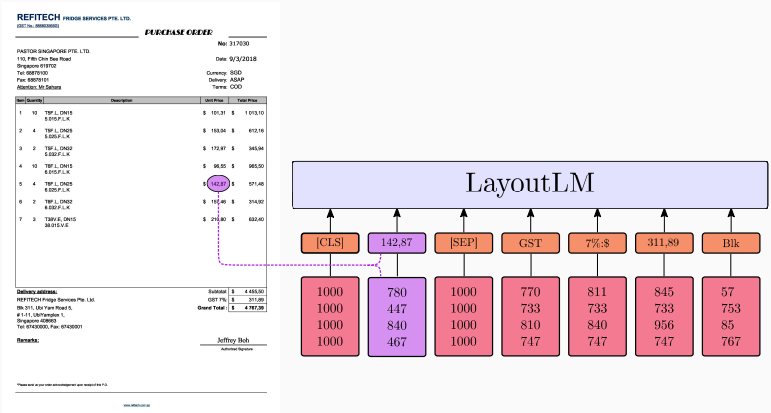


Figure 5: A figure in the document is randomly selected.

Numeric Ordering

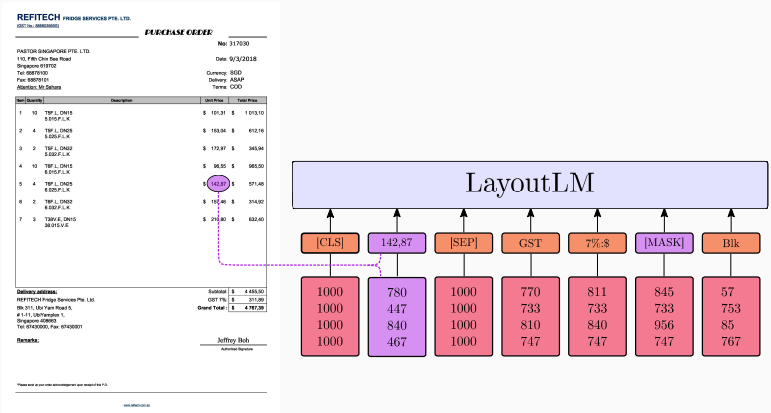


Figure 5: Some tokens and / or positions are randomly masked or replaced.

Numeric Ordering

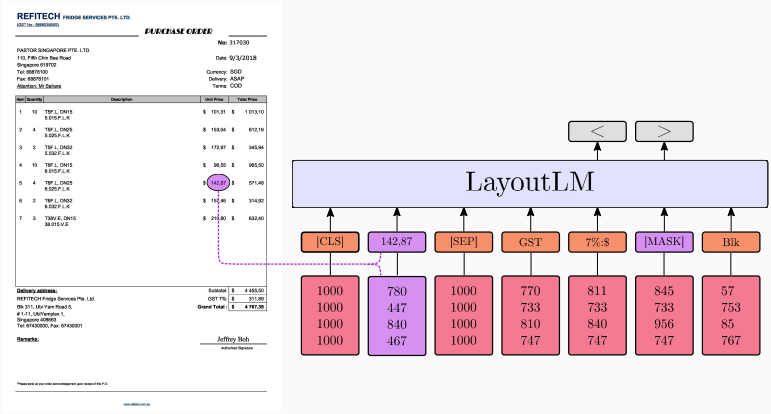
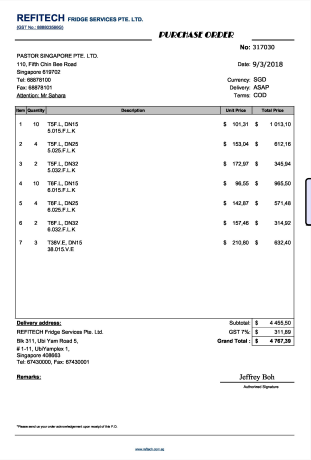


Figure 5: Model is pre-trained on the prediction of relative order between each number and the chosen one.

Layout Inclusion



REFITECH FRIDGE SERVICES PTE. LTD.

GST No.: 888833999G

PURCHASE ORDER

No: 3170030

PASTOR SINGAPORE PTE. LTD.
110, Fifth Chin Bee Road
Singapore 619702
Tel: 68978100
Fax: 68978101
Attention: Mr Sahara

Date: 9/3/2018

Currency: SGD
Delivery: ASAP
Terms: COD

Item	Quantity	Description	Unit Price	Total Price
1	10	TSF L. DN1S 5.015.F.L.K	$ 101,31	$ 1 013,10
2	4	TSF L. DN2S 5.025.F.L.K	$ 153,04	$ 612,16
3	2	TSF L. DN32 5.032.F.L.K	$ 172,97	$ 345,94
4	10	TSF L. DN1S 6.015.F.L.K	$ 96,55	$ 965,50
5	4	TSF L. DN2S 6.025.F.L.K	$ 142,87	$ 571,48
6	2	TSF 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, LSI/Tampies 1,
Singapore 408863
Tel: 67400000, 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

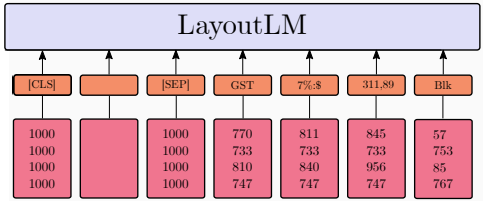


Figure 6: Layout Inclusion (LI) teaches the relative position of tokens. Only part of the sequence is represented.

Layout Inclusion

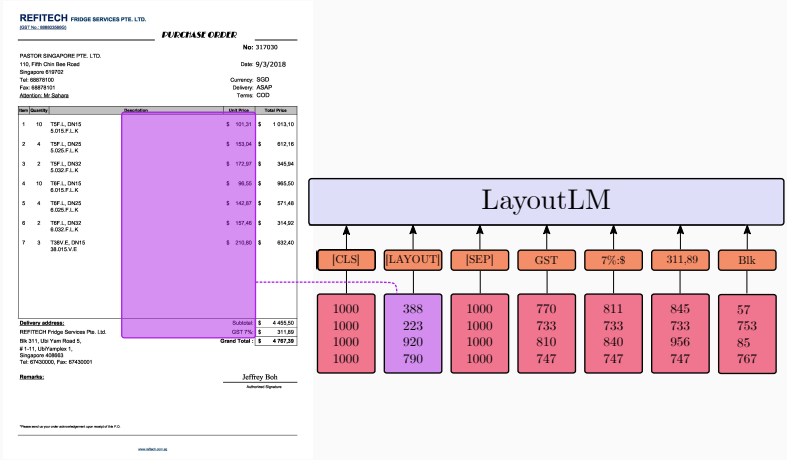


Figure 6: A rectangle zone is randomly chosen inside the document's boudaries. It is represented by a *[LAYOUT]* token.

Layout Inclusion

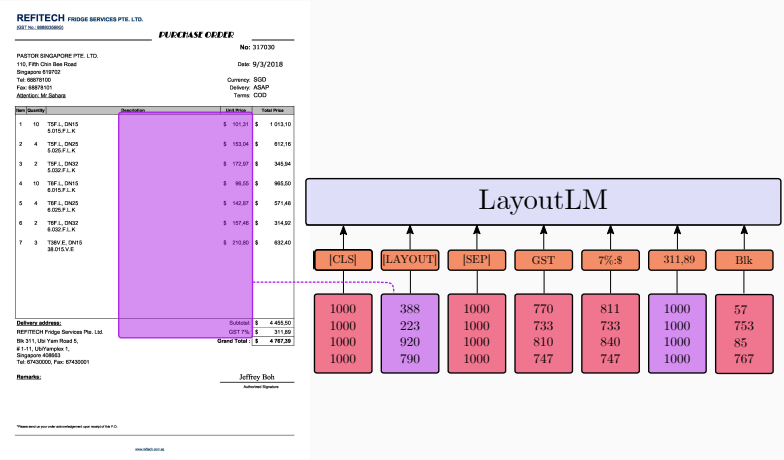


Figure 6: Some tokens and / or positions are randomly masked or replaced.

Layout Inclusion

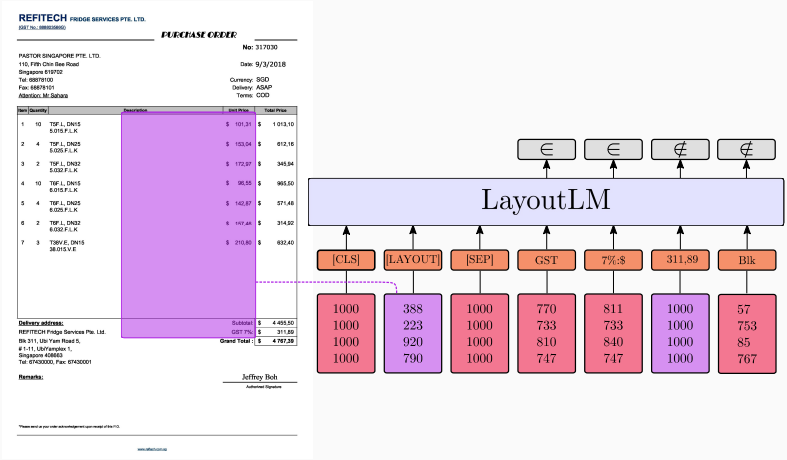


Figure 6: Model predicts for each token if its center is inside or outside the random rectangle.

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

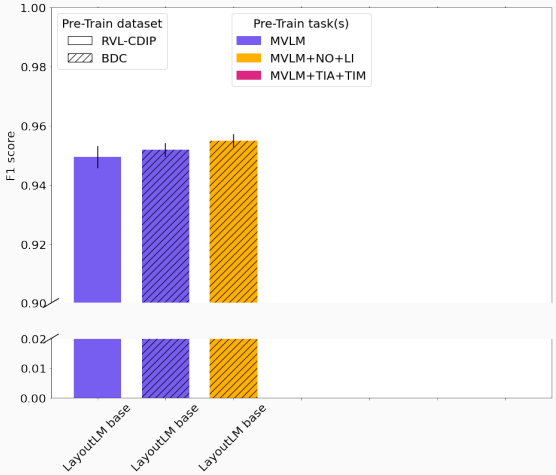


Figure 7: Results on SROIE fine-tuning. Both BDC and the new pre-training tasks improve model performance.

Results

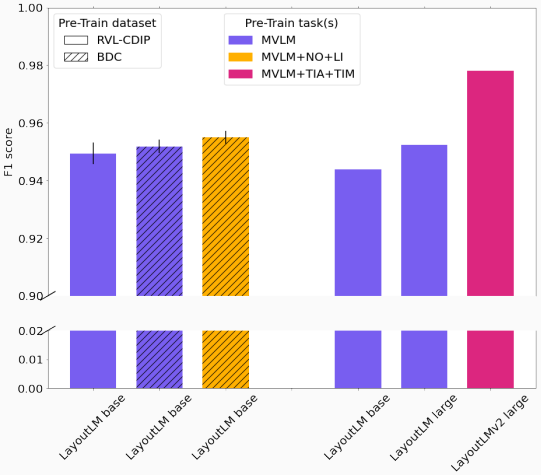


Figure 7: Our pre-trained models (left) outperform original LayoutLM [5, 6] base and large models (right).

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 Jérémy 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`