ViBERTgrid BiLSTM-CRF: Multimodal Key Information Extraction from Unstructured Financial Documents

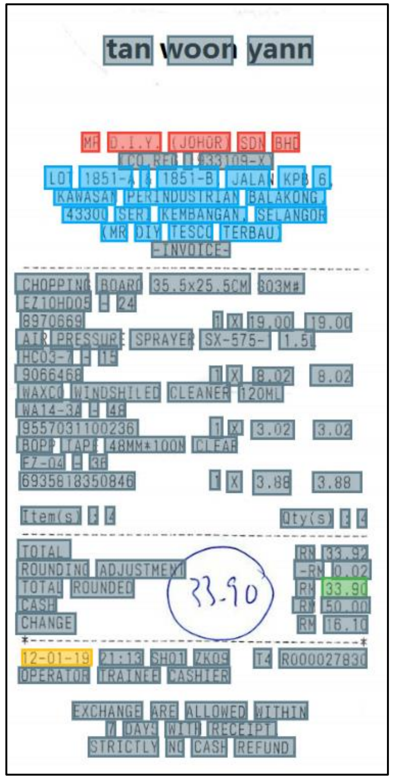
Mehmet Yasin AKPINAR

YapıKredi
Teknoloji

Problem Statement

- **Key Information Extraction (KIE)** from document images is the automated process of retrieving relevant data presented visually in the document.

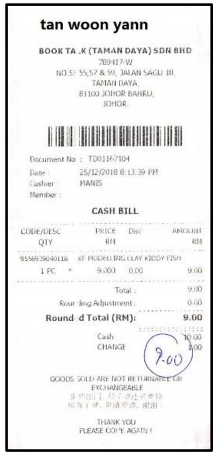
- **Optical Character Recognition (OCR)** tools are used to extract the text in the document.



```
tan woon yann

MR D.I.Y. (JOHOR) SDN BHD
(CO.REG :933109-X)
LOT 1851-A & 1851-B, JALAN KPB 6,
KAWASAN PERINDUSTRIAN BALAKONG,
43300 SERI KEMBANGAN, SELANGOR
(MR DIY TESCO TERBAU)
-INVOICE-
------------------------------------------------
CHOPPING BOARD 35.5x25.5CM 803M#
EZ10HD05 - 24
8970669                    1 X 19.00    19.00
AIR PRESSURE SPRAYER SX-575-1 1.5L
HC03-7 - 15
9066468                    1 X  8.02     8.02
WAXCO WINDSHILED CLEANER 120ML
WA14-3A - 48
9557031100236              1 X  3.02     3.02
BOPP TAPE 48MM*100M CLEAR
E7-04 - 36
6935818350846              1 X  3.88     3.88

Item(s) : 4                      Qty(s) : 4
------------------------------------------------
TOTAL                               RM 33.92
ROUNDING ADJUSTMENT                -RM  0.02
TOTAL ROUNDED          33.90        RM 33.90
CASH                                RM 50.00
CHANGE                              RM 16.10
*----------------------------------------------*
12-01-19 21:13 SH01 ZK09     T4 R000027830
OPERATOR TRAINEE CASHIER

EXCHANGE ARE ALLOWED WITHIN
7 DAYS WITH RECEIPT.
STRICTLY NO CASH REFUND
```

Problem Statement

- Using the OCR text, many studies model the KIE task as a **sequence tagging** problem and **solve using NER** (Yu et. al, 2021).



tan woon yann

BOOK TA .K (TAMAN DAYA) SDN BHD
789417-W
NO.5: 55,57 & 59, JALAN SAGU 18,
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.

Document No : TD01167104
Date : 25/12/2018 8:13:39 PM
Cashier : MANIS
Member :

CASH BILL

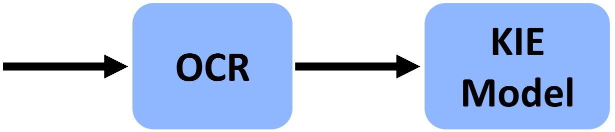
CODE/DESC PRICE Disc AMOUNT
QTY RM RM
9556939040116 KF MODELLING CLAY KIDDY FISH
1 PC * 9.000 0.00 9.00

Total : 9.00
Rour ding Adjustment : 0.00
Round d Total (RM): 9.00
Cash 10.00
CHANGE 1.00

9.00

GOODS SOLD ARE NOT RETURNABLE OR
EXCHANGEABLE
[illegible]
[illegible]

THANK YOU
PLEASE COME AGAIN !



- **Multimodal** approaches combine the OCR output with visual information of the document image and position information of the tokens.

tan woon yann

BOOK TA .K (TAMAN DAYA) SDN BHD
789417-W
NO.5: 55,57 & 59, JALAN SAGU 18,
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.

Document No : TD01167104
Date : 25/12/2018 8:13:39 PM
Cashier : MANIS
Member :

CASH BILL

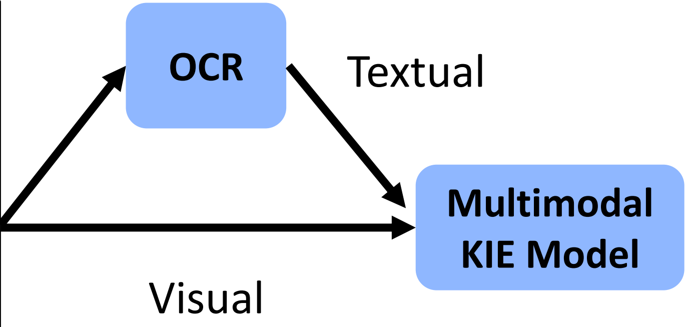
CODE/DESC PRICE Disc AMOUNT
QTY RM RM
9556939040116 KF MODELLING CLAY KIDDY FISH
1 PC * 9.000 0.00 9.00

Total : 9.00
Rour ding Adjustment : 0.00
Round d Total (RM): 9.00
Cash 10.00
CHANGE 1.00

9.00

GOODS SOLD ARE NOT RETURNABLE OR
EXCHANGEABLE
[illegible]
[illegible]

THANK YOU
PLEASE COME AGAIN !



Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4363– 4370. IEEE (2021)

Document Types

Fixed-form / structured documents

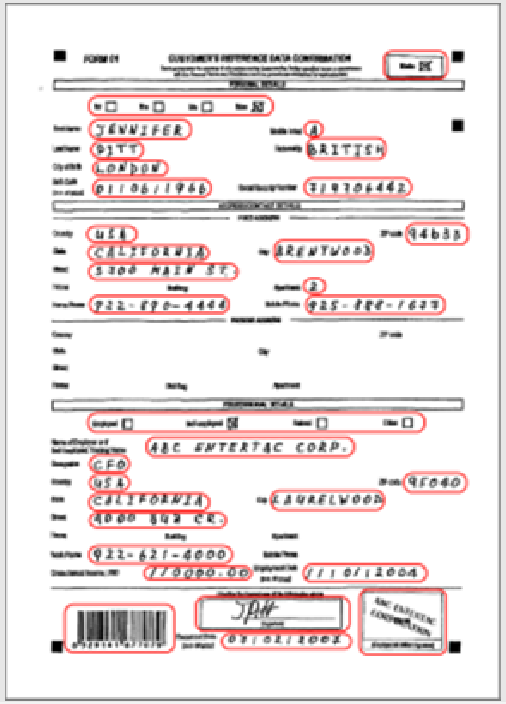
- Surveys
- Questionnaires
- Tests
- Claim forms

Semi-structured documents

- Invoices
- Purchase orders
- Bills of lading
- EOBs

Unstructured documents

- Contracts
- Letters
- Articles
- Notes



FORM 01

CUSTOMER'S REFERENCE DATA CONFIRMATION

JENNIFER

PITT

LONDON

01/06/1966

BRITISH

319706442

USA

CALIFORNIA

BRENTWOOD

94633

5300 MAIN ST.

922-870-4444

925-888-1637

ABC ENTERTAC CORP.

CFO

USA

CALIFORNIA

LAURELWOOD

95040

4000 SUR CR.

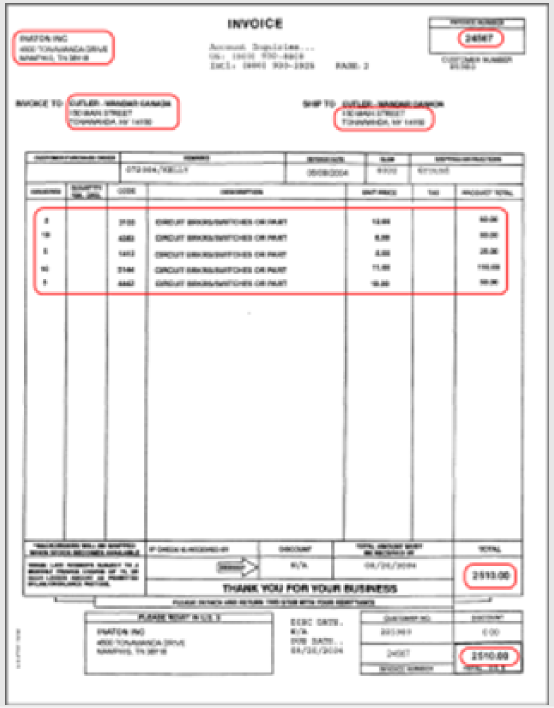
922-621-4000

170000.00

11/01/2004

07/02/2002

ABC ENTERTAC CORPORATION



INVOICE

24587

INVOICE TO: CUTLER - HAMMER CANADA

SHIP TO: CUTLER - HAMMER CANADA

CIRCUIT BRKRS/SWITCHES OR PART

CIRCUIT BRKRS/SWITCHES OR PART

CIRCUIT BRKRS/SWITCHES OR PART

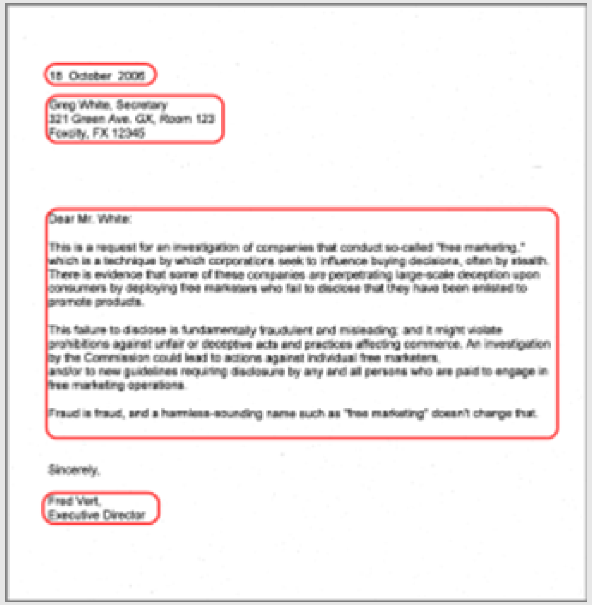
CIRCUIT BRKRS/SWITCHES OR PART

CIRCUIT BRKRS/SWITCHES OR PART

2510.00

THANK YOU FOR YOUR BUSINESS

2510.00



18 October 2006

Greg White, Secretary
321 Green Ave. GX, Room 123
Foxcity, FX 12345

Dear Mr. White:

This is a request for an investigation of companies that conduct so-called "free marketing," which is a technique by which corporations seek to influence buying decisions, often by stealth. There is evidence that some of these companies are perpetrating large-scale deception upon consumers by deploying free marketers who fail to disclose that they have been enlisted to promote products.

This failure to disclose is fundamentally fraudulent and misleading; and it might violate prohibitions against unfair or deceptive acts and practices affecting commerce. An investigation by the Commission could lead to actions against individual free marketers, and/or to new guidelines requiring disclosure by any and all persons who are paid to engage in free marketing operations.

Fraud is fraud, and a harmless-sounding name such as "free marketing" doesn't change that.

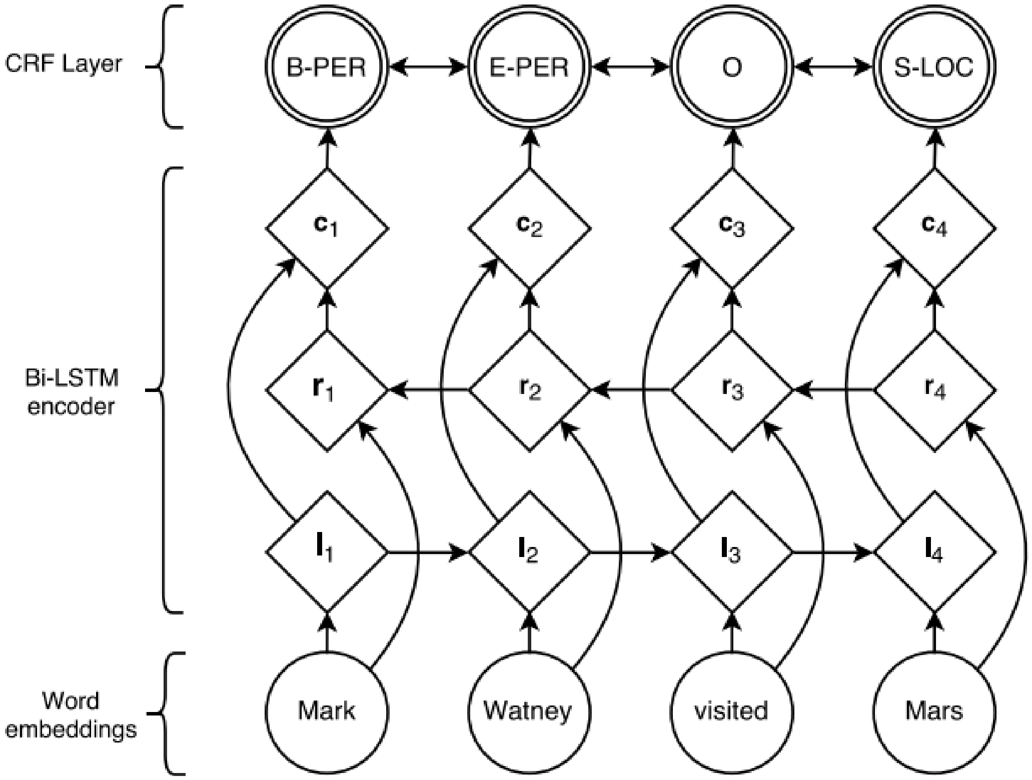
Sincerely,

Fred Vert,
Executive Director

Previous Methods and Limitations

BiLSTM-CRF (Huang et. al, 2015)

- Pros:
  - Effective in sequence tagging and NER tasks.
  - Uses both past and future tokens.
  - Capture sentence level meaning.
- Cons:
  - Uses only textual information.
  - Unable to utilize layout or visual information directly.



Works best on unstructured documents.

Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

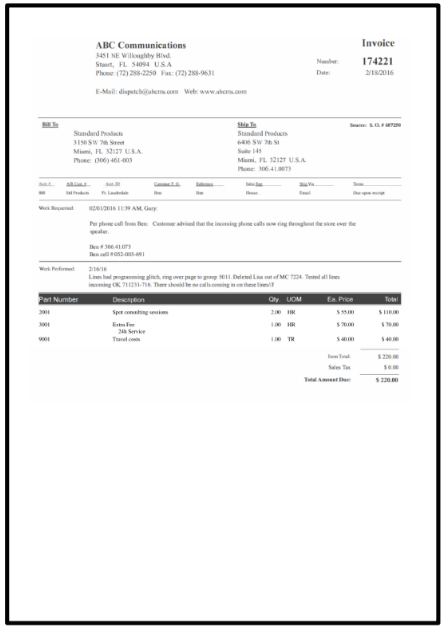
Previous Methods and Limitations

Chargrid (Katti et. al, 2018)

- Pros:
  - Converts a page into 2D grid of chars.
  - Encodes spatial features.
  - Better in structured documents.

- Cons:
  - Using only char level information is not sufficient for understanding the semantics.
  - Thus, not as effective in unstructured documents as in structured ones.

Raw data



ABC Communications
3451 NE Willoughby Blvd.
Stuart, FL 54094 U.S.A
Phone: (72) 288-2250 Fax: (72) 288-9631

E-Mail: dispatch@abcms.com Web: www.abcms.com

Invoice
Number: **174221**
Date: 2/18/2016

Bill To
Standard Products
3150 SW 7th Street
Miami, FL 32127 U.S.A.
Phone: (306) 461-003

Ship To
Standard Products
6406 SW 7th St
Suite 145
Miami, FL 32127 U.S.A.
Phone: 306.41.0073

Source: S. O. # 687250

Acct. #	A/R Cust. #	Acct. ID	Customer P. O.	Reference	Sales Rep	Ship Via	Terms
800	Std Products	Ft. Lauderdale	Ben	Ben	House .	Email	Due upon receipt

Work Requested: 02/01/2016 11:39 AM, Gary:

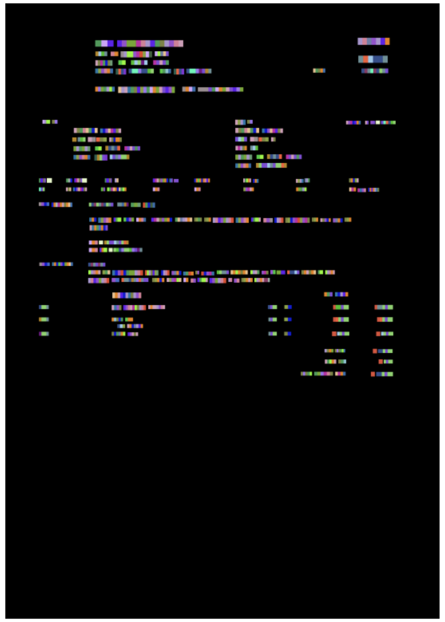
Per phone call from Ben: Customer advised that the incoming phone calls now ring throughout the store over the speaker.

Ben # 306.41.073
Ben cell # 052-005-691

Work Performed: 2/16/16
Lines had programming glitch, ring over page to group 5011. Deleted Lisa out of MC 7224. Tested all lines incoming OK 711231-716. There should be no calls coming in on these lines//J

Part Number	Description	Qty.	UOM	Ea. Price	Total
2001	Spot consulting sessions	2.00	HR	$ 55.00	$ 110.00
3001	Extra Fee 24h Service	1.00	HR	$ 70.00	$ 70.00
9001	Travel costs	1.00	TR	$ 40.00	$ 40.00
				Item Total:	$ 220.00
				Sales Tax	$ 0.00
				Total Amount Due:	**$ 220.00**

Chargrid



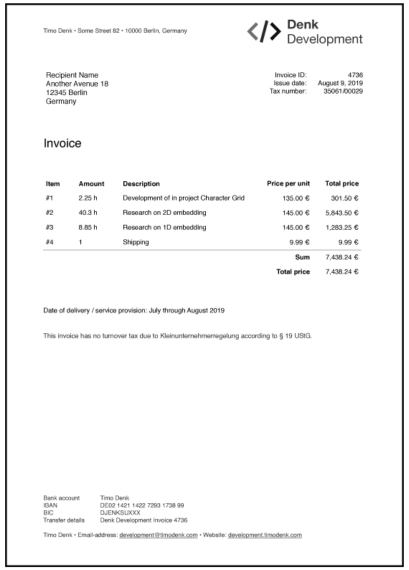
Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., H¨ohne, J., Faddoul, J.B.: Chargrid: Towards understanding 2d documents. arXiv preprint arXiv:1809.08799 (2018)

Previous Methods and Limitations

BERTgrid (Denk et. al, 2019)

- Pros:
  - Converts a page into 2D grid of contextualized word embeddings obtained from BERT.
  - Possible to understand the semantics of document.

- Cons:
  - BERT language model is frozen during training.
  - Image of the document page is not directly utilized.



Timo Denk • Some Street 82 • 10000 Berlin, Germany

Denk
Development

Recipient Name
Another Avenue 18
12345 Berlin
Germany

Invoice ID:	4736
Issue date:	August 9, 2019
Tax number:	35061/00029

Invoice

Item	Amount	Description	Price per unit	Total price
#1	2.25 h	Development of in project Character Grid	135.00 €	301.50 €
#2	40.3 h	Research on 2D embedding	145.00 €	5,843.50 €
#3	8.85 h	Research on 1D embedding	145.00 €	1,283.25 €
#4	1	Shipping	9.99 €	9.99 €
			Sum	7,438.24 €
			Total price	7,438.24 €

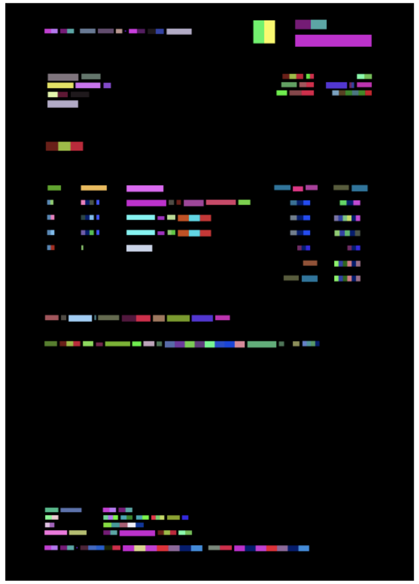
Date of delivery / service provision: July through August 2019

This invoice has no turnover tax due to Kleinunternehmerregelung according to § 19 UStG.

Bank account	Timo Denk
IBAN	DE02 1421 1422 7293 1738 99
BIC	DJENKSUXXX
Transfer details	Denk Development Invoice 4736

Timo Denk • Email-address: development@timodenk.com • Website: development.timodenk.com

Raw image



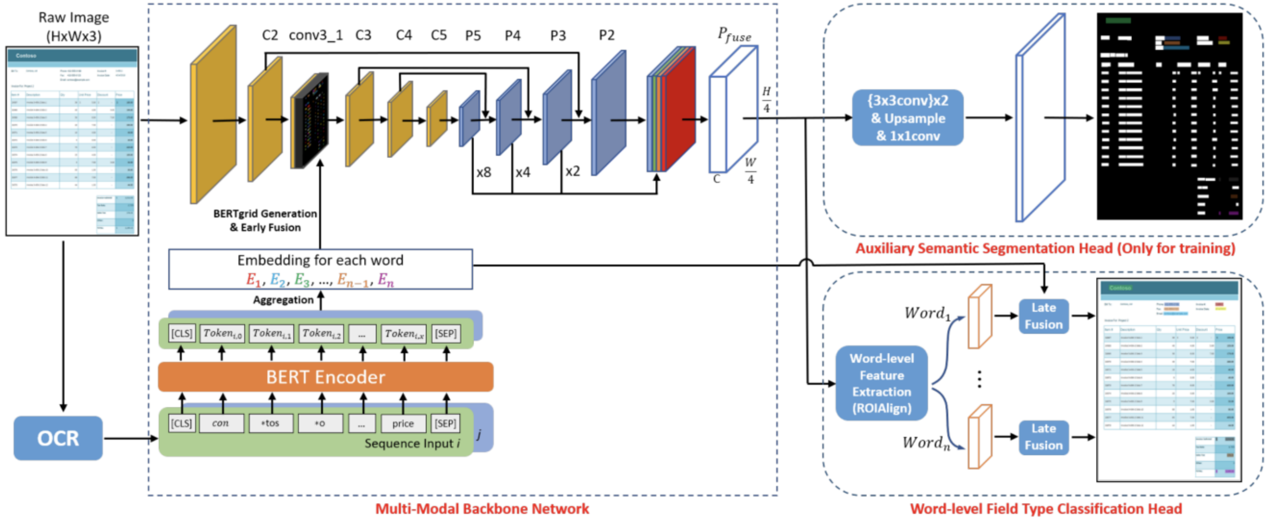
BERTgrid representation

Denk, T.I., Reisswig, C.: Bertgrid: Contextualized embedding for 2d document representation and understanding. arXiv preprint arXiv:1909.04948 (2019)

Previous Methods and Limitations

ViBERTgrid (Lin et. al, 2021)

- Pros:
  - Incorporates BERTgrid with a CNN to process the raw image directly.
  - Jointly trains BERT and CNN.



- Cons from our observations:
  - Could not outperform a pure textual model (BiLSTM-CRF on BERT emb.) on unstructured money transfer order documents.

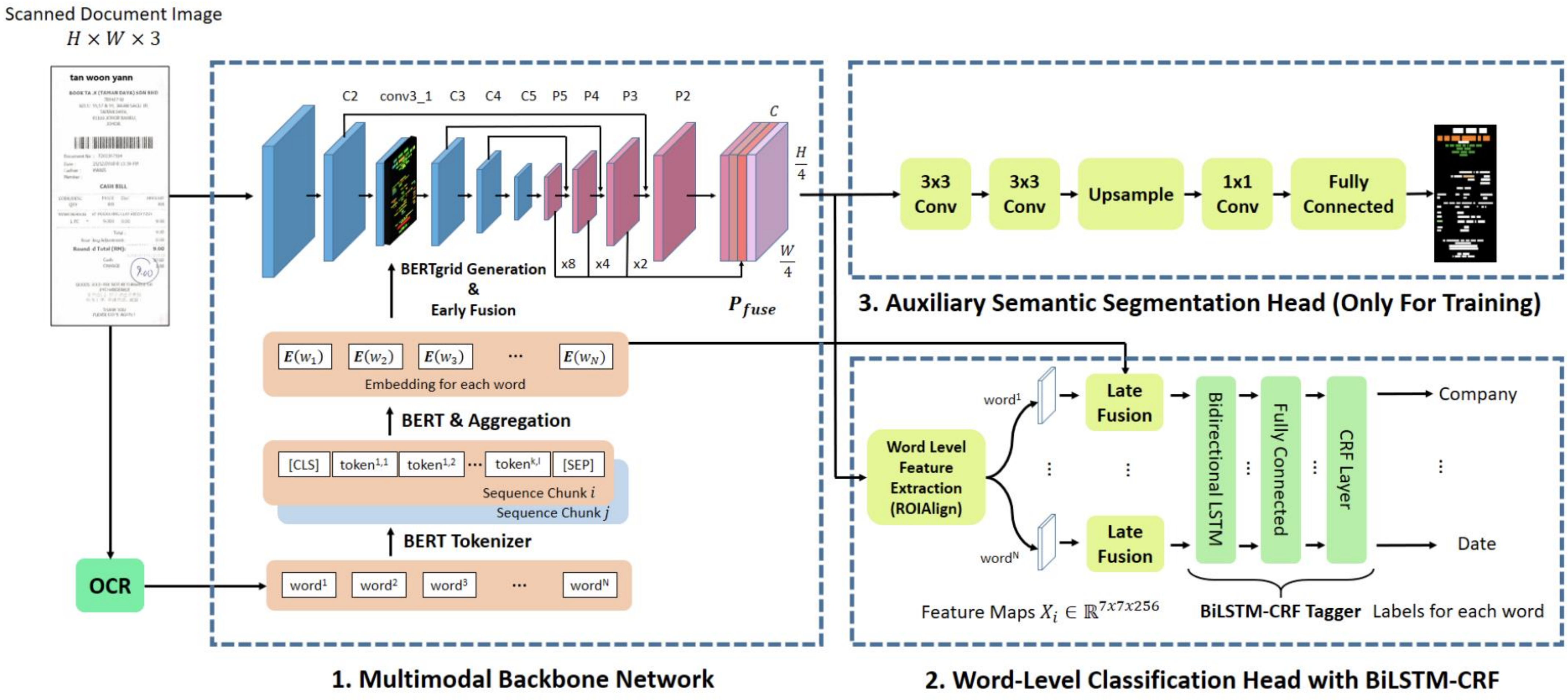
Works best on structured documents.

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. In: International Conference on Document Analysis and Recognition. pp. 548–563. Springer (2021)

Idea

What if we combine the best performing models?

Proposed Approach: ViBERTgrid BiLSTM-CRF

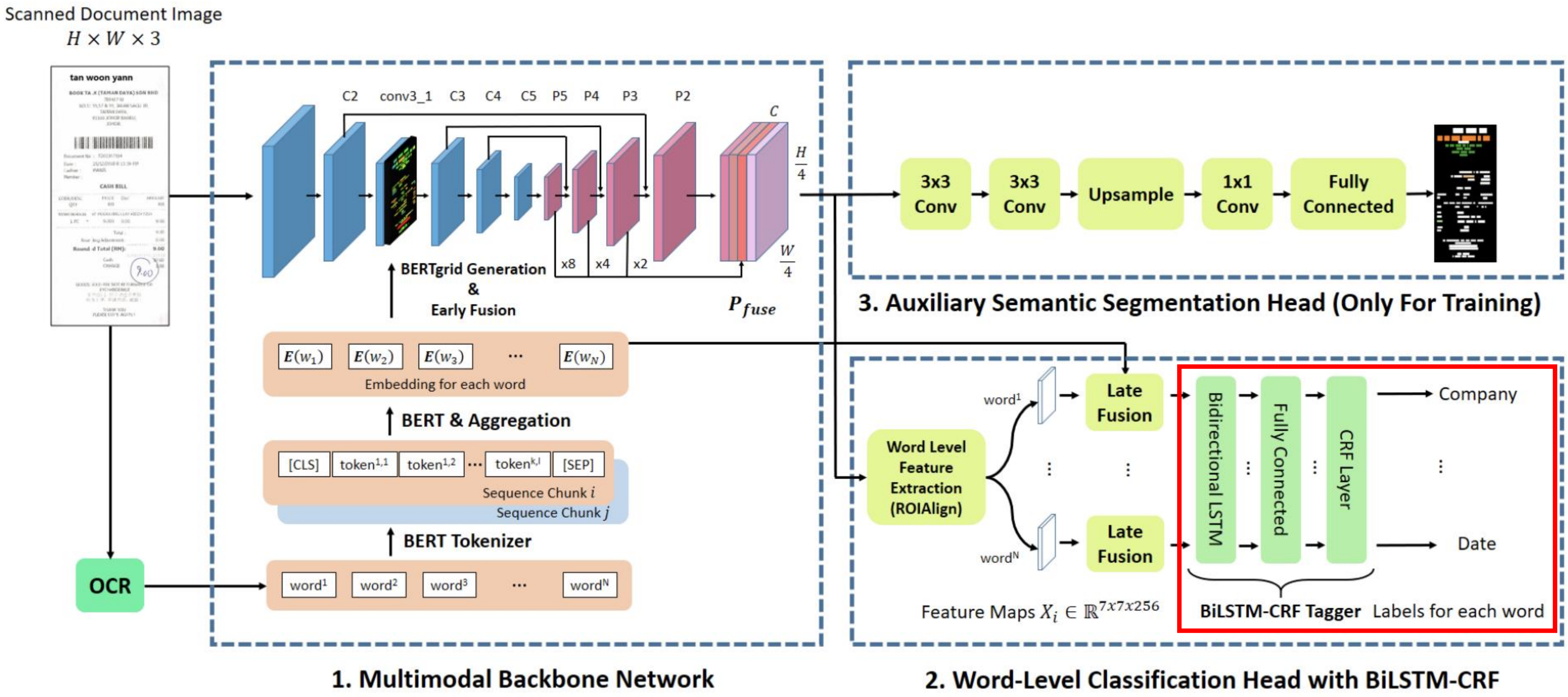


1. Multimodal Backbone Network

2. Word-Level Classification Head with BiLSTM-CRF

3. Auxiliary Semantic Segmentation Head (Only For Training)

Proposed Approach: ViBERTgrid BiLSTM-CRF

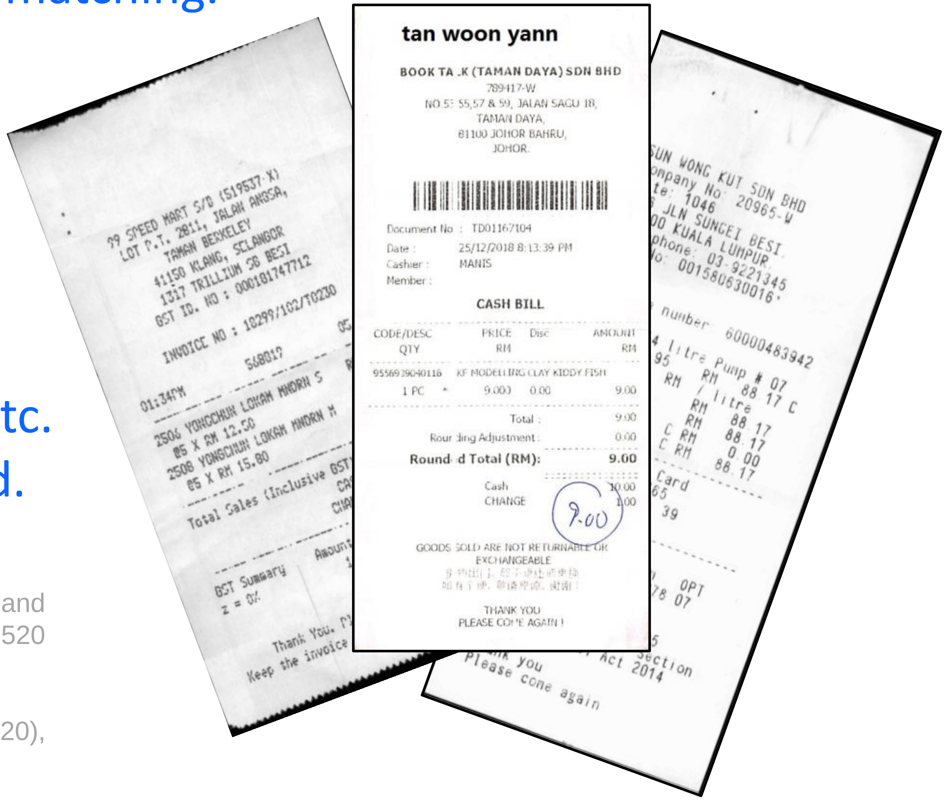


Scanned Document Image
$H \times W \times 3$
tan woon yann
OCR
C2
conv3_1
C3
C4
C5
P5
P4
P3
P2
C
$\frac{H}{4}$
$\frac{W}{4}$
x8
x4
x2
BERTgrid Generation & Early Fusion
P_{fuse}
$E(w_1)$ $E(w_2)$ $E(w_3)$ ··· $E(w_N)$
Embedding for each word
BERT & Aggregation
[CLS] $\text{token}^{1,1}$ $\text{token}^{1,2}$ ··· $\text{token}^{k,l}$ [SEP]
Sequence Chunk i
Sequence Chunk j
BERT Tokenizer
word^1 word^2 word^3 ··· word^N
1. Multimodal Backbone Network
3x3 Conv
3x3 Conv
Upsample
1x1 Conv
Fully Connected
3. Auxiliary Semantic Segmentation Head (Only For Training)
Word Level Feature Extraction (ROIAlign)
word^1
word^N
Late Fusion
Late Fusion
Feature Maps $X_i \in \mathbb{R}^{7x7x256}$
Bidirectional LSTM
Fully Connected
CRF Layer
Company
Date
BiLSTM-CRF Tagger
Labels for each word
2. Word-Level Classification Head with BiLSTM-CRF

Datasets

SROIE

- ICDAR SROIE dataset[1]: 973 receipts (626 training, 347 testing samples).
- Semi-structured documents.
- Four entity types: company, date, address, total.
- Dataset presents key information fields and OCR output separately.
- Previous studies[2,3] used text-based matching but results in poor matching.
- We manually annotated the entire dataset on the word-level.
- We publicly release the word-level annotations of SROIE dataset for use in multimodal transformers.
  (https://github.com/YKT-NLP/ICDAR-2019-SROIE-Token-Level-Annotations)
- Evaluation on test set is still problematic due to discrepancies between OCR output and key information fields, e.g., mismatched punctuation, extra or missing white spaces, typos etc.
- Some of these errors have been documented and manually fixed.



tan woon yann

BOOK TA .K (TAMAN DAYA) SDN BHD
789417-W
NO.5: 55,57 & 59, JALAN SAGU 18,
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.

Document No : TD01167104
Date : 25/12/2018 8:13:39 PM
Cashier : MANIS
Member :

CASH BILL

CODE/DESC QTY	PRICE RM	Disc	AMOUNT RM
9556939040116	KF MODELLING CLAY KIDDY FISH		
1 PC *	9.000	0.00	9.00

Total : 9.00
Rour ding Adjustment : 0.00
Round d Total (RM): 9.00

Cash 10.00
CHANGE 1.00

9.00

GOODS SOLD ARE NOT RETURNABLE OR EXCHANGEABLE
[illegible]

THANK YOU
PLEASE CO!'E AGAIN !

[1] 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 (2019). https://doi.org/10.1109/ICDAR.2019.00244

[2] Lin, Z.: Vibertgrid pytorch (2021), https://github.com/ZeningLin/ ViBERTgrid-PyTorch

[3] Delplace, A.: Chargrid model : Extraction of meaningful instances from document images (2020), https://github.com/antoinedelplace/Chargrid

Datasets

Transactional Documents

- In-house dataset consists of unstructured Turkish money transfer order documents, introduced by Oral et. al (2022).
- Has two sets: **U**nstructured **T**ransactional **D**ocuments (**UTD**) and **U**nstructured **M**ulti-**T**ransaction **D**ocuments (**UMTD**).
- UTD has 3500 documents (2500 for training, 400 validation, 600 testing).
- UMTD has 1154 documents (954 for training, 200 testing).
- Within the UMTD test set, 54 out of 200 documents have tabular-like layouts (**TLL**), the rest has non-tabular-like (**noTLL**) documents.
- We used the same splits as in Oral et. al (2022) for consistency.

Oral, B., Eryiğit, G.: Fusion of visual representations for multimodal information extraction from unstructured transactional documents. International Journal on Document Analysis and Recognition (IJDAR) pp. 1–19 (2022)

Experiments & Results

SROIE

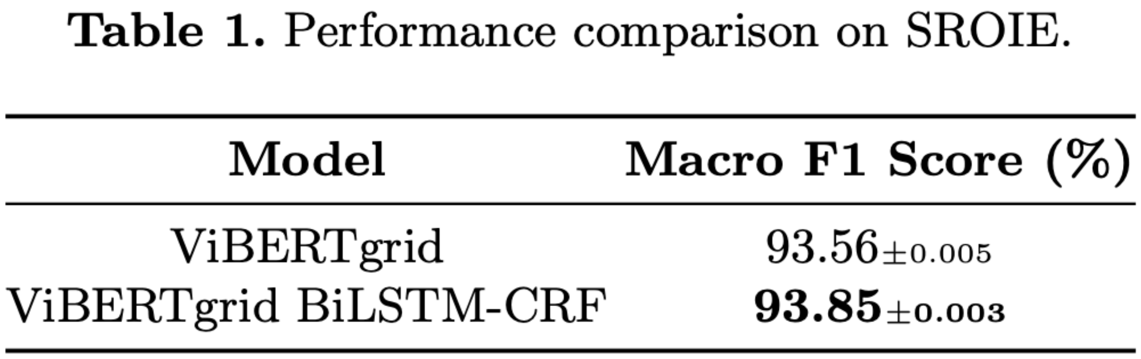
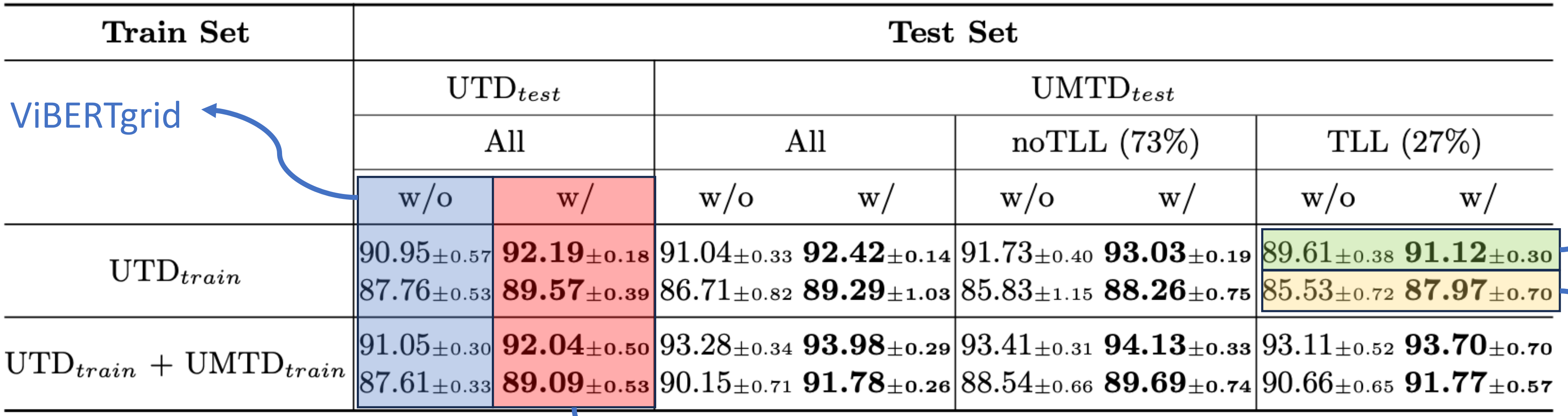


Table 1. Performance comparison on SROIE.

Model	**Macro F1 Score (%)**
ViBERTgrid	$93.56_{\pm 0.005}$
ViBERTgrid BiLSTM-CRF	$\mathbf{93.85}_{\pm \mathbf{0.003}}$

Experiments & Results

Transactional Documents



Train Set	Test Set							
	UTD$_{test}$		UMTD$_{test}$					
	All		All		noTLL (73%)		TLL (27%)	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
UTD$_{train}$	90.95$\pm$0.57	**92.19**$\pm$**0.18**	91.04$\pm$0.33	**92.42**$\pm$**0.14**	91.73$\pm$0.40	**93.03**$\pm$**0.19**	89.61$\pm$0.38	**91.12**$\pm$**0.30**
	87.76$\pm$0.53	**89.57**$\pm$**0.39**	86.71$\pm$0.82	**89.29**$\pm$**1.03**	85.83$\pm$1.15	**88.26**$\pm$**0.75**	85.53$\pm$0.72	**87.97**$\pm$**0.70**
UTD$_{train}$ + UMTD$_{train}$	91.05$\pm$0.30	**92.04**$\pm$**0.50**	93.28$\pm$0.34	**93.98**$\pm$**0.29**	93.41$\pm$0.31	**94.13**$\pm$**0.33**	93.11$\pm$0.52	**93.70**$\pm$**0.70**
	87.61$\pm$0.33	**89.09**$\pm$**0.53**	90.15$\pm$0.71	**91.78**$\pm$**0.26**	88.54$\pm$0.66	**89.69**$\pm$**0.74**	90.66$\pm$0.65	**91.77**$\pm$**0.57**

vanilla ViBERTgrid

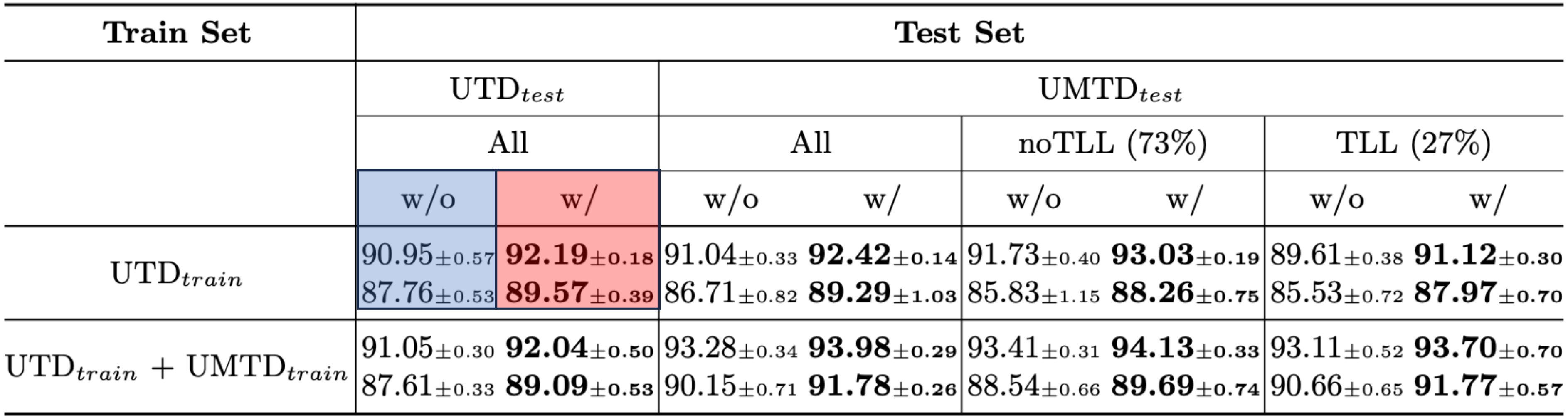
micro F1

macro F1

ViBERTgrid with BiLSTM-CRF (ours)

Experiments & Results

Transactional Documents

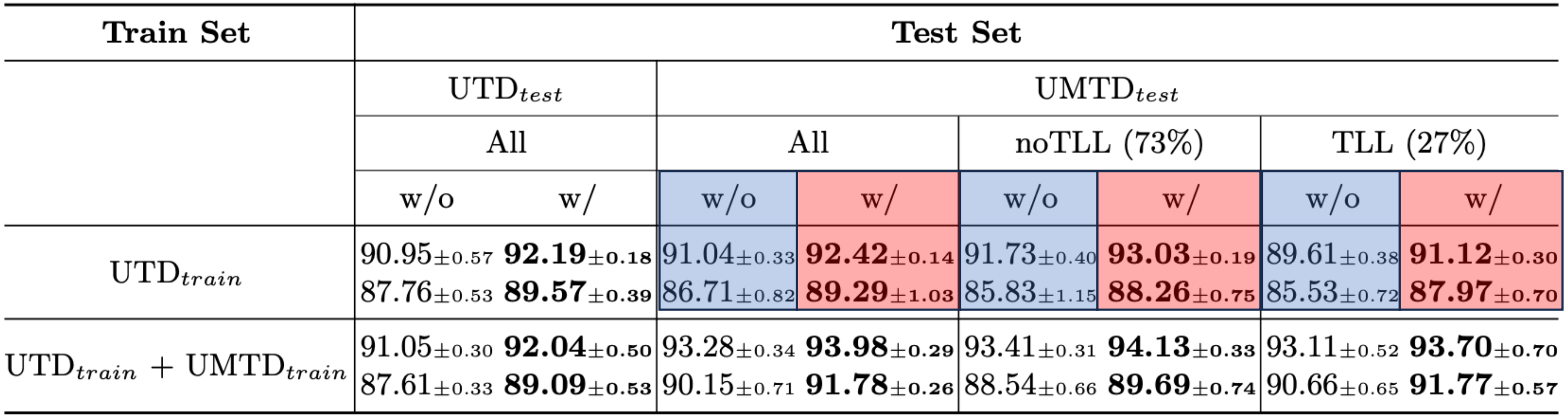


Train Set	**Test Set**							
	UTD_{test}		UMTD_{test}					
	All		All		noTLL (73%)		TLL (27%)	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
UTD_{train}	$90.95_{\pm0.57}$ $87.76_{\pm0.53}$	$\mathbf{92.19}_{\pm\mathbf{0.18}}$ $\mathbf{89.57}_{\pm\mathbf{0.39}}$	$91.04_{\pm0.33}$ $86.71_{\pm0.82}$	$\mathbf{92.42}_{\pm\mathbf{0.14}}$ $\mathbf{89.29}_{\pm\mathbf{1.03}}$	$91.73_{\pm0.40}$ $85.83_{\pm1.15}$	$\mathbf{93.03}_{\pm\mathbf{0.19}}$ $\mathbf{88.26}_{\pm\mathbf{0.75}}$	$89.61_{\pm0.38}$ $85.53_{\pm0.72}$	$\mathbf{91.12}_{\pm\mathbf{0.30}}$ $\mathbf{87.97}_{\pm\mathbf{0.70}}$
$\text{UTD}_{train} + \text{UMTD}_{train}$	$91.05_{\pm0.30}$ $87.61_{\pm0.33}$	$\mathbf{92.04}_{\pm\mathbf{0.50}}$ $\mathbf{89.09}_{\pm\mathbf{0.53}}$	$93.28_{\pm0.34}$ $90.15_{\pm0.71}$	$\mathbf{93.98}_{\pm\mathbf{0.29}}$ $\mathbf{91.78}_{\pm\mathbf{0.26}}$	$93.41_{\pm0.31}$ $88.54_{\pm0.66}$	$\mathbf{94.13}_{\pm\mathbf{0.33}}$ $\mathbf{89.69}_{\pm\mathbf{0.74}}$	$93.11_{\pm0.52}$ $90.66_{\pm0.65}$	$\mathbf{93.70}_{\pm\mathbf{0.70}}$ $\mathbf{91.77}_{\pm\mathbf{0.57}}$

Oral et. al (2022) obtained a micro and macro F1 score of **91.48** and **88.45** on UTD dataset using a textual only BiLSTM-CRF model pretrained on BERT embeddings.

Experiments & Results

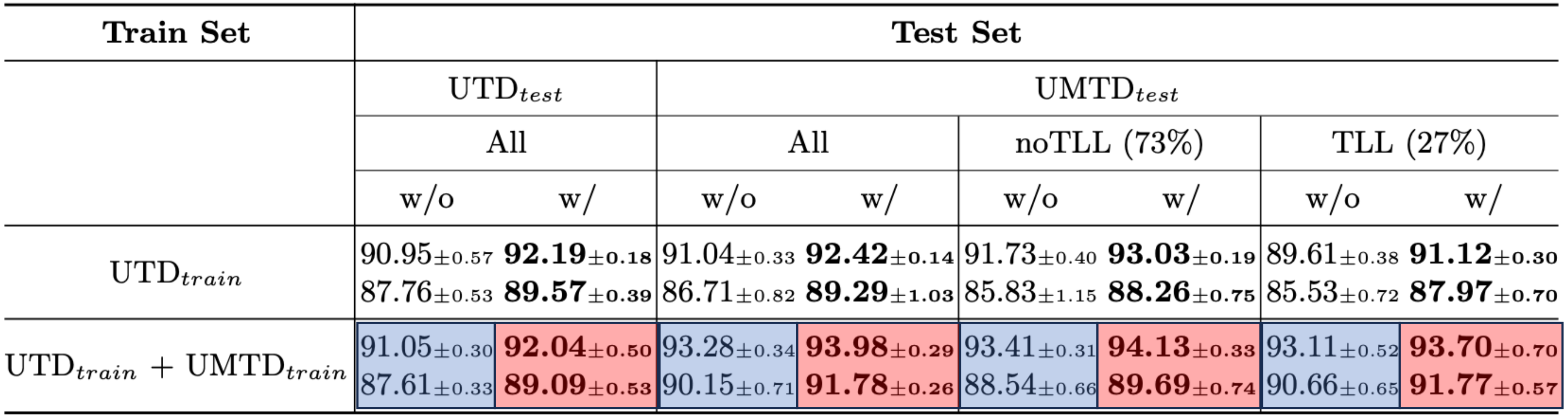
Transactional Documents



Train Set	Test Set							
	UTD$_{test}$		UMTD$_{test}$					
	All		All		noTLL (73%)		TLL (27%)	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
UTD$_{train}$	90.95$_{\pm0.57}$ 87.76$_{\pm0.53}$	**92.19**$_{\pm\mathbf{0.18}}$ **89.57**$_{\pm\mathbf{0.39}}$	91.04$_{\pm0.33}$ 86.71$_{\pm0.82}$	**92.42**$_{\pm\mathbf{0.14}}$ **89.29**$_{\pm\mathbf{1.03}}$	91.73$_{\pm0.40}$ 85.83$_{\pm1.15}$	**93.03**$_{\pm\mathbf{0.19}}$ **88.26**$_{\pm\mathbf{0.75}}$	89.61$_{\pm0.38}$ 85.53$_{\pm0.72}$	**91.12**$_{\pm\mathbf{0.30}}$ **87.97**$_{\pm\mathbf{0.70}}$
UTD$_{train}$ + UMTD$_{train}$	91.05$_{\pm0.30}$ 87.61$_{\pm0.33}$	**92.04**$_{\pm\mathbf{0.50}}$ **89.09**$_{\pm\mathbf{0.53}}$	93.28$_{\pm0.34}$ 90.15$_{\pm0.71}$	**93.98**$_{\pm\mathbf{0.29}}$ **91.78**$_{\pm\mathbf{0.26}}$	93.41$_{\pm0.31}$ 88.54$_{\pm0.66}$	**94.13**$_{\pm\mathbf{0.33}}$ **89.69**$_{\pm\mathbf{0.74}}$	93.11$_{\pm0.52}$ 90.66$_{\pm0.65}$	**93.70**$_{\pm\mathbf{0.70}}$ **91.77**$_{\pm\mathbf{0.57}}$

Experiments & Results

Transactional Documents



Train Set	Test Set							
	UTD$_{test}$		UMTD$_{test}$					
	All		All		noTLL (73%)		TLL (27%)	
	w/o	w/	w/o	w/	w/o	w/	w/o	w/
UTD$_{train}$	90.95$_{\pm 0.57}$ 87.76$_{\pm 0.53}$	**92.19**$_{\pm 0.18}$ **89.57**$_{\pm 0.39}$	91.04$_{\pm 0.33}$ 86.71$_{\pm 0.82}$	**92.42**$_{\pm 0.14}$ **89.29**$_{\pm 1.03}$	91.73$_{\pm 0.40}$ 85.83$_{\pm 1.15}$	**93.03**$_{\pm 0.19}$ **88.26**$_{\pm 0.75}$	89.61$_{\pm 0.38}$ 85.53$_{\pm 0.72}$	**91.12**$_{\pm 0.30}$ **87.97**$_{\pm 0.70}$
UTD$_{train}$ + UMTD$_{train}$	91.05$_{\pm 0.30}$ 87.61$_{\pm 0.33}$	**92.04**$_{\pm 0.50}$ **89.09**$_{\pm 0.53}$	93.28$_{\pm 0.34}$ 90.15$_{\pm 0.71}$	**93.98**$_{\pm 0.29}$ **91.78**$_{\pm 0.26}$	93.41$_{\pm 0.31}$ 88.54$_{\pm 0.66}$	**94.13**$_{\pm 0.33}$ **89.69**$_{\pm 0.74}$	93.11$_{\pm 0.52}$ 90.66$_{\pm 0.65}$	**93.70**$_{\pm 0.70}$ **91.77**$_{\pm 0.57}$

Conclusion

- We focused on the impact of using a multimodal transformer (i.e., ViBERTgrid previously explored on semistructured documents) on the NER task from **unstructured financial documents**.
- The initial results showed that the original ViBERTgrid has a **negative impact on unstructured** documents compared to a pure textual baseline.
- We presented an approach to enhance the performance of ViBERTgrid on unstructured documents by **extending it with a BiLSTM-CRF layer.**
- As a result, our proposed **ViBERTgrid BiLSTM-CRF** model demonstrated a significant improvement in performance (**up to 2 percentage points**) on unstructured documents, while maintaining its performance on semi-structured documents, in the **domain of financial and banking documents**.
- As an additional contribution, we publicly released token-level annotations for the SROIE dataset to pave the way for its use in multimodal sequence labelling models.

Q & A ?