## Information Extraction from Text Regions with Complex Tabular Structure

Kaixuang Zhang, Zejiang Shen, Jie Zhou, & Melissa Dell

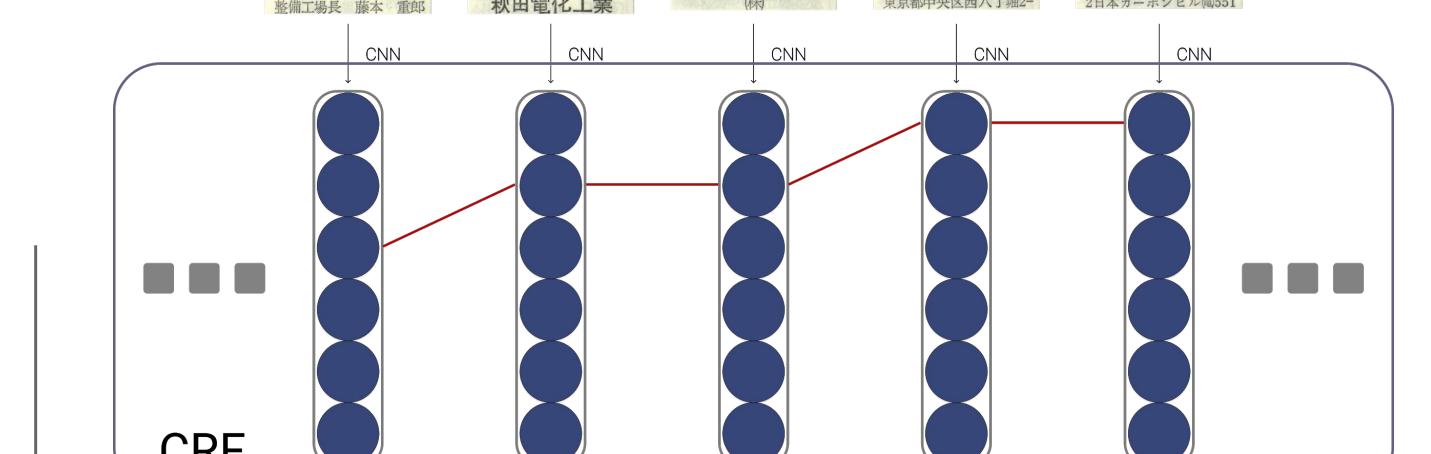
There has been an increasing interest in document analysis in recent years, with significant progress in automatic layout segmentation of document images into text and non-text regions. However, efforts to parse the information in text region have been limited. The challenge of parsing text region comes from their potentially complex structure: information within text regions may not simply read from left to right or top to down.

This challenge is particularly severe in historical documents, as in the past document formats were much less standardized than today. Vast amounts of historical data that could shed light on important economic issues remain locked in hard copy due to prohibitive curation costs.



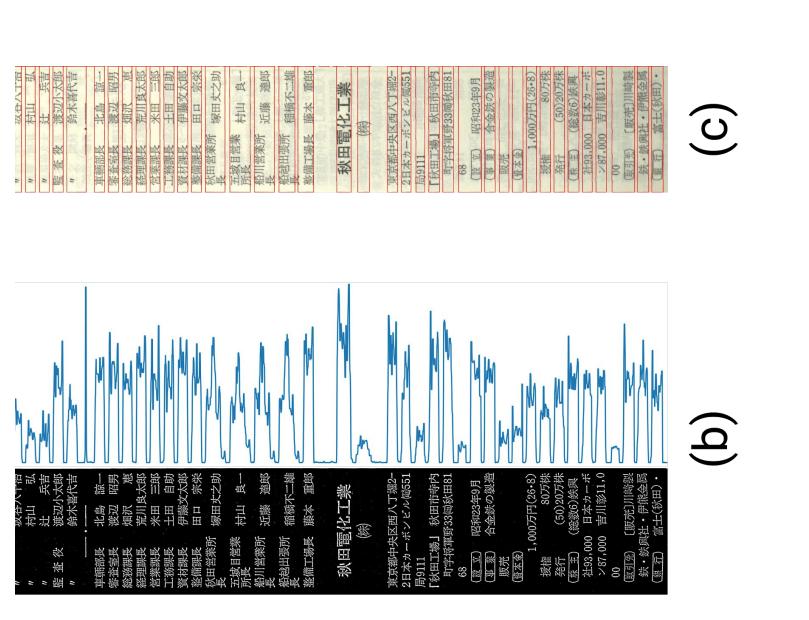


The Institute for Quantitative Social Science



To study the problem of historical document analysis, we collected around 1,000 scanned images of Personnel Record 1956 (PR1956), which is a Japanese economic document containing a rich variety of economic, personnel, and financial variables for around 15,000 Japanese firms in year 1956.

Each text region has five columns and each column has a number of rows. Therefore, the correct way to interpret the text region is to parse it by column from left to right, and parse each column by row from top to bottom. For each row, we classify them into six categories.



The purpose of segmentation is to

CRF

Row classification is vitally important for extracting structured information from the tabular structured data. And we utlized CNN and CRF for classification.

- CNN is used to extract features from row images.
- CRF is applied to model the dependency between rows.

**Data** We randomly choose four text regions from PR1956 and manually label their row images (1407 row images in total) with correspondent class. Three text regions are used as the training (85%) and validation (15%) set, the other one text region is used as a test set. For the training and validation set, image data augmentation techniques are applied to artificially expand the amount of data.

**CNN** A MobileNet(v1) is trained with our training data for 40 epochs. We experimented with different augmentation methods and hyperparameters, and the model with the highest validation accuracy is selected as the model for inference.

Segmentation

acquire the basic unit for document analysis.

Taking row segmentation for example, we binarize the column image and count the number of pixels horizontally. Then we estimate row bounding boxes according to the one dimension signal in (b).

(a)

和歌山工場

里見

会計部

工場長 松山

工器具付 井上

部 長 大場

##### 村上

監売員 潮尾

8長村 復谷

動势課長 山角

血感課長 中井

使課長 原

裂油部

高岡 弥太郎

長 田島

代明 田井

代理 阿部

雪田調長 河合

即 6 長 植木

工務課長

带具信 村上

工事課具 前谷

維造 精製環長

浅野

田辺

举 同代理 家永

枪務

長 加納

土橋

販売部

《實際長前 国幹 "

電 長 古田 又彦 <u>原類執計</u>

- 古田 武夫 检察品的

高荷課長 山中 順一部

策務部

部 哲 小林

調査課長 松岡

5 長 倍

企業課長 并桁

预造漂兵 守口

6兵·夏 西技調課 国分

日長付 野島

口林副長 结谷

長 高稽

長 平田

F 员 古賀

場長 大橋

海南新 坂井田

起 務 部

日县提

清水工場

会計部

經理異長 小倉 真一部

开具付 椰木

**唐澤長** 別府

製造部

同代明 竹田 正明

技術部

品質管理部

6 村 竹沢 泰夫

大阪事務所

府 村 長谷川 恵之 蓋型問題

太田 敬一 種加潤長

再究蜜長 長尾 福太郎 目代 电 成田

價統計 學驗 育後 動力課長 開弘

其他说動員價

長期貸债

资本金

(現刊語)

■総理具 宮井 新太郎 デイアラビア・インドキシ

再評價積立会

常期純利益

利政济偿金

经延期定

同系合計借入金 1.280.000

價格經動準備金 334,60

機總利佔剩餘金 1,452,1

30年12月 104第2.47275円

30年6月 5億4-692万円 30年12月 8億3,298万円

弘明 (111至) 30年6月 2185分

(建東西) 1.916

「資水工場」 部間禁密原用袖 節可損1900億三九月60~6

「和歌山工術」和歌山靜海軍

即初為可能而1000億其為55

大阪事務所」 北張梅田町47

5 新設時ビル第三大員91~2

東京福宁代田区大手町

(夏夏) 昭和17年5月 (夏夏) 農業業務の製造版

(選手査) 4,000万円 (昭和29年12月)

速·全國都這府發展消費業

(油TE) 会國總定件業務和 維延稳定

尾上 哲之助

弘海 2-8團和田倉(20)2746

· 3522 · 4987-8

設備 320万株

長 西沢

協乱速

純一 専 赛 渡辺

目代理 仁木 敏久 常 喜 中山 幸三郎

前川長三郎

退職給與引言金 230.920 紅苔賞長 富永

1.157.000 工爆長· 池川 63.692 總將主任 池川

製造主任 佐藤

銀行預金

請 貯 会

受取手刑

賢掛金

穩卸資產-

任任者

常鮮品

委託员

假佛金

假 捲 点

10 70 4

生命保障金

有形固定资度

精築物

機械發展

水脉驱激具

工具器具備品

建設價額定

预形固定选择

施設利用權

投票

出资金

有效認知

前纬费用

支持手形

質損金

宋 禄 位

假受会

成動員值

貢債・資本

3.157.000

(<u>雪夏</u>) 30年6月 77億2,409万円 貴虐資產

30年12月 2期5分

Emission score comes from the trained CNN, transition score CRF is manually setup (1: row pair can appear; 0: row pair cannot appear).

Method	CNN	CNN+CRF
Accuracy	95.6%	96.8%

了错落了握了 潜北质川东西2

東亜バルブ朝

曼 太田 常太郎

红 欣一

大坪 释:

松福

藍直夜 松水 潤清

劳西国兵 计

15.001 設計課具 土井

63.555 作業課具 小泉

電 務 福

刻造信

部 指重 古田 百日

: 務 格原 夏州

(制改 木下

頁 兵庫県尼崎市水燈字創

寬閉 林1113篇大版四大层

→ 紅田市 本原 赤幸 (軍東市) 6

石石石石 富永

7 作借対照表

水學所做其

**新桃田市** 

池川

000 工場長。 692 趙指士任

製造主任

的现代的 建合立 534

11.101 (192 至) 30年6月 2回5会

· 次· 頁 【過水工場】 節型盤處原面給

○ 長田島 康弘 西初島町龍斎1000度東島35 母格金

新始技術 高岡 小太郎 新設設ビル第三大目91-2 調査系

太一 東亜農薬(#)

(19-30) 昭和17年5月 (正常) 陸軍原州の副造縦

出 E) 今月該宣告堂協和 展示数学

· 今回起派府县所济县军 前接費用

精次 金 計 1四部 (音 昭) 20年4月 30年4月

福祉調長 宮井 新太郎 デイアラビア・インドキシ

目代理 田井 久時 0~7 「大阪東勝平」 全原利的町0

Yar merger propriet       Name merger propris       Name merger propriet       N	<ul> <li>韓國惠特·律村 完成</li> <li>英方面形 究成</li> <li>新方面形 賀集 範木 友次席</li> <li>勤劳面形 賀集 節七 次方面 費</li> <li>前方面形 賀集 節子</li> <li>前方面形 竹木 方次席</li> <li>前方面形 賀集 節子</li> <li>前方面形 竹木 方次席</li> <li>前方面形 賀集 節子</li> <li>前方面形 竹木 市</li> <li>「一方子</li> &lt;</ul>		購買課具 狩野 方 新管課具 大島 仁 (死亡!!!) 6月・12月 住借対照表(30・12來) 克 差 千円 液動資産 3.571.895 保行冠金現金 111.558 受取構定 124.429 泉 油 655.796 未留原油 1.123.894 國品申製品 872.758 每 四 1.740 貯 截 品 639.290 共偏旋動資産 43.427 
日代理 浅田 弦雄 重時課題 藤田 容嗣 法 規 部 二 務 部 二 本課題 為 持 二 教 部 二 本課題 為 特 二 本課題 為 特 二 本課題 為 行 二 教 部 二 本課題 為 特 二 本課題 為 行 二 本課題 人 在 一 本課題 一 10.2555.803 一 工 時 二 本課題 人 在 一 本課題 一 10.2555.803 一 代 四 一 (本) 7.60 二 本課題 一 10.2555.803 一 代 四 一 (本) 7.60 二 一 (本) 7.60	法规部 都 長 古谷 障男 機械撤定 102,236 常業課長 尾上 祐二 (原来10) 300	法規部 電長古谷 南明 振動で 102,235 営業課長 配上 祐二 (22単1) 300 電長 土田 敏想 工事課長 高封 上 其他資産 40,705 大阪香業 内の 安 「大阪香菜所」 北国堂為上町	關定管理: 5.638.145 說話開定 102.236 其他要證 40.706 合 計 10.255.803 資值: 查末 補助資值 2.405.632 支揚手形得入金 879.530 未揚點定 460.582 計税引含金 930.515

長卵 中山 率三郎

1 工事具: 木原 寿幸 OEEEE 6

小田原工場

横浜工場

貸借対照表 (22・9末)

17 11

氯古國南非2681個議府律5

「橫濱工場」 香北溪川和町25

東亜バルブ制

《康则尼崎市水景字》

長太田常太郎太田工業適會

(昭和28年4月)

創立合資の後昭和15年4月

大国党太郎36.500 大島星

专20.000 西華產業12.00

a 具 谷田 友差

古田

大坪

松浦

2 聖查役 松本 潤油

常務部

医 長頭 柿原 夏歸

次 問題 杠 欣一直

起音課兵 太田 象離

勞難調長 辻 史郎

製造部

部 兵廠 吉田 正臣

" 卿 大坪 軽之

#理課時 片间

5 次 長ी 松浦.

15.901 設計課長 土井 秀

63.556 作業課長 小泉

山下 真吾

杠 欣一点

以 取雜役 木下 寅市

2 社 長 太田 常太郎

10-502 常 喜 相原 夏雄

東印パルプ製造機設立

林1113週大阪四八周5

千円 (征用) 大正11年3月現社

10.664 (単面) 高温高能パルプ報

4.325 (資志至) 2.100万円

151.877 登棚 8754.000株

26.923 股行 4772.000株

1.957 當15.900

7.550 (直正) 三连点率40.000

97.209 西南韩蒙12.008 日興證書

500118011 + 41

ation

 $\mathbf{C}$ 

Siff

Class

nt

rime

Exp

$\mathbf{O}$
IJ
11
D
•

	■人間長 伊藤 漢支整理 → 1-22	1	1 and a state of the	刊	さ信	10 JKB	開定 初定会	460-5 930-5	00 小槽角索 82 所長 15	111		100	内町7-26	北海道小樽の	
	課長 八勝	赤田	1	间联山		其後	规制费用	134.2	20 mr stins	製 造 data	- chi	14	加して用し お何の決さ	间京川都小田 581儒派府律5	H I
	販売能	5	工場長				放動員引 (会計冊)	7 1.260.0	73	tripe-		0-	-1		
	西县林	I	工事具付				員債	31.0		田原口	C28)	1000	CC根」 I	能再用和何2	5
	原實課長 南	国幹		里見		- <b>FR</b> 32	精定		9. 製造調展	木原	. Ale		ALARTIT		1
100	唐荷躍長 山中	颜一朗	a second second	会計			統與引言 種動準備	金 230.9		富永		12			
Sec.	業務部		10 14			· 費 · · ·	*	5.973.6		黄派工		1	重バ	ルブ朝	
	部 長 古田	双言		100		100.04	K≜.	3,159.00	10 工爆展。	池川		-			
100	部 信 小林	正夫	開展	野村	欣-		價稅立會 明信会	465.10				町 長坊	观尼崎)	专术堂学岛	
	* 古田	武夫	NUMBER OF	村上	1	そ 機能	利益刺自	1.452.63	1	佐藤	nt?	り 淋目	13個)大信	民四八月58	
	調花漂長 松岡	正幹	and the second sec	上山	新設		純和估 計	832.97	7 貸借支	分照表	(29 - 93	0 61-4		and the second second	
	常務課長 別府	能次	营造具	濒尾				10.255-80	一 把放花用	5 16			D 大正	11年3月現記	
	製造部	MER		地 路		3040	5.13 7	7億2:609358	5 会会你が		108.45	7 <u>H</u>	(田常太阳)	<b>太田工業商会</b>	
	商员佃	-	图 長	加納	修-	- (MERLE		4間2,472.255	9 現	\$	27		「日前の便」	E finska H	1
100	a manual start	聖之	部具有	複谷	ħ	3044	5.71 53	4-692万円	銀行書 調 取		10.66	4 OE	D高副	真見パルプ製	
S. Same	A	治郎		土橋	<b>弘</b> 原	30-90	2月 8]	15.298万円	5111		4.32	3 值景	D 2.100		
	and the second second	忠失	動例源與	山角			304	6月 21月53 12月 21月55	要掛	4	5.10			(2)[9] 第128年4月)	1
	19104	正明	MIL THE	富井	in the second	OR SLA	) (仕)	「北國・中ガ	N NT	15	151.87		8754.00	OFF	
100	技術部		12824A		新太郎	913	7787	· 1> F+5		許	28.92		4752.00	0.000 (0140 - 000)	
	過發調課 国分	-*	同代理	中井	IE	CHETE	) 當士	- 10.00	1 E	野	66	5 大田	音大郎36。	500 大岛滨	
	原長付 野岛 3		保安課品	家水	定植	CHERT	1.91	1.1.1	任挺	品	4.02	920	.000 29.4	國產業12.000	
		第久二		NI.	- 海男	[清水工]		电频率原因数 三大局-0~6	1 市县	品	1,95		900	C II Million	1
	La	虎夫		2 油 1				to What has a state of the stat			6	2	各田	友拍	
	品質管理部		10 10	田島	康弘		时政治	000開発鳥-55	假得会		13.478		1 太田		1
	部 员 高橋	武弘	目代理想的政治	田井	久晴	0~7	1991	加加和小	價語	企	10.502	12 1	5 和原	夏雄	
1	即 付 竹沢	泰夫	IN 16	高同	弥太郎			三大局列-2	后 近 報 页 走	2	4.049	1025	古田		
	* 浜田	敏衛	目代理	阿部	瑞男	1000			/生命保/	協会	1.7.02			IEIZ	1
	大阪事務所		製油管理	大高	太一	का :	西粤	東側	有形因是自	-	69.117	1	112.1	寅市	
	所 長 平田		同代用	八木	姚次	-	IL AR	>#Con		ity .	65-077		山下	真否	
	所具付 櫃木	寬方	-	浅野	AP.DIL	東口に作る	千代田	区大手町	精装车		1,328	-	杠	欣一郎	
	研究所		推測課具	河合	瓜海			20)2746	· 機械與3 本種的		20,712 3,068	-	大坪	桥之	
	所 员 古賀	意造	则有消兵	中山	茂	• 3522			工具数		2.710	教査者	松調		
1	所 付 長谷川	恵之	初州助教 調兵	田辺	和夫	CEED			土 九		2.207	- 10 D	and the	洞治	
	* 太田	111-	操放調具	通山		CIDD	用双规	間の知道戦	建设使某 例可进定者	記書	3.921 3.190	-	常務日	and the second	
1	呼樂蜜具 長尾 褶			成田	忠友 義人	STERED.	4.0001	ET.	歌 4	2	2,333	第二 長田	all the	夏雄	
	清水工場			務會			- OBI	四年(2月)	推迟 FIII	140	852	次 長田	and a	欣一郎	1
-	工場員 大樹	三朝	西 長	19-1-	34	授權 : 附行	120.25世紀		四 我 金	1	850 650	新L的28.55		和增	
	会計部	100			治	CRLED	全國講	官商党如何	有权消费	1	200			-2	
- 8	5 長 堤			村上	an in	(理、全)	和政治	算刑消费常	維定助定 前陸費用		700	劳胜部兵		史郎	
	<b>新聞課長</b> 小会 直	des 1		時谷	時越	制度逐			盘 計		700 258.315		製造會	5	
10	期他計 你前			武川	an l	社員	西沢	文地	発信	• 費本		四 長御		正臣	
10	時用納 arran ret			同弘	胞一	市商	渡辺	嘉二	何動員信 支持手用		143.915 58.164		THE REAL		
1	ERE NOCHLITT		120101-0-	仁木		常ち	中山	幸三郎	買撥金		15,901	- 00	大坪	释之	
	載 務 部	1	「「「「「「」」」」	前川	長三郎		尾上	智之助	<b>火持会</b> 假受会			設計課具	土井	現	
				100		-			N X E	-	SV21200		小泉	清二	
-								-				-	1.1.1	- Line	
															TAXABLE IN

scanned image with detected text region

column segmentation

row segmentation

秀 耳代星 仁末 國久 常 喜 中川 幸三郎 星世金

和歐山工場

工場長 松山

幹 型見 図 会計部 算 表 大場

**ビガガ長 村上** 

又意 原有統計 野村

龄列进兵

京務課長 中井

堂 目代照 微永

保安課長 原

值新 目代母 阿部 瑞明

整洁管理 大島

感觉观共

精整调线

1 工作課具 開谷

育後 喬力漂具 調弘

輕短調長 小倉 真一部 工種調長 武川

吉田 武夫 植海岛的

阿在課長 松岡

合國課長 井村

同代明 析田

西拉德國 国分

大阪事務所

row classification

共植見越負位

其他被動員戲

同素合趾借入金 1.280.000 | 長期貨位 31.000 |

退職結與引含金 230.920

價格提號即個金 334.600

第 末 5,973,640 1 費 末 金 3,159,000 1 再評價額並金 63,692 1 利金階層金 465,100

嚴國利益副營金 1,452,871

公期純利益 832.977 合 21 - 10.255.803

30年6月 77位2,609万P

30年12月 10402,472万円

30年6月 5億4-692万円 30年12月 6億5-298万円

(NET) 30446月 2間5分 304412月 2開5分

度可定的 〔任人〕未開・サウ

デイアクビア・インドネシ

(夏夏) お士・異銀 (夏夏) 1,916 「別水江湖」 韓国新地区間

和歌山工場上和歌山等称

郡初乌町政治1000與民乌

「大戰以臨帝」 北盔得田町4

新阪師ビル同三大員?1~3

東亜農薬㈱

東京都千代田区大手叩

(第三) 昭和1745月 (三) 段繁要報の発売時

(中国語) 4.000万円 (中国29412月

资行 80万株 CMEED 全时期订合无路4

社 長 西沢 文雄

母 裔 渡辺 嘉二

常 將 中山 韓三郎

理·全國部造府部長的資業

援抓 32075株

協机应

· 3522 · 4987-8

2-800和田倉(20)2746 Address

梯时祖1900周三大局60

Variable

Company

Value

134,230

773