Information Extraction from Text Regions with Complex Tabular Structure

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

HARVARD
Faculty of Arts and Sciences
DEPARTMENT OF ECONOMICS

IQSS
The Institute for Quantitative Social Science

Motivation

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This challenge is particularly severe in historical documents, as in the past document formats were much less standardized than today.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.

Dataset

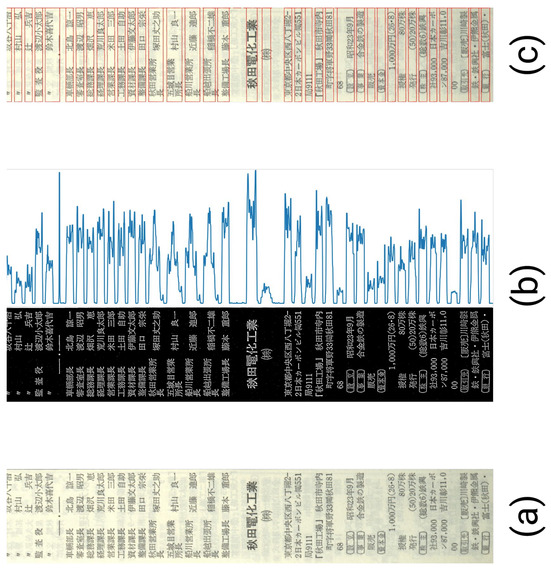
To study the problem of historical document analysis, we collected around 1,000 scanned images of Personnel Record 1956 (PR1956), which is a Japanese economic document containing a rich variety of economic, personnel, and financial variables for around 15,000 Japanese firms in year 1956.

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.

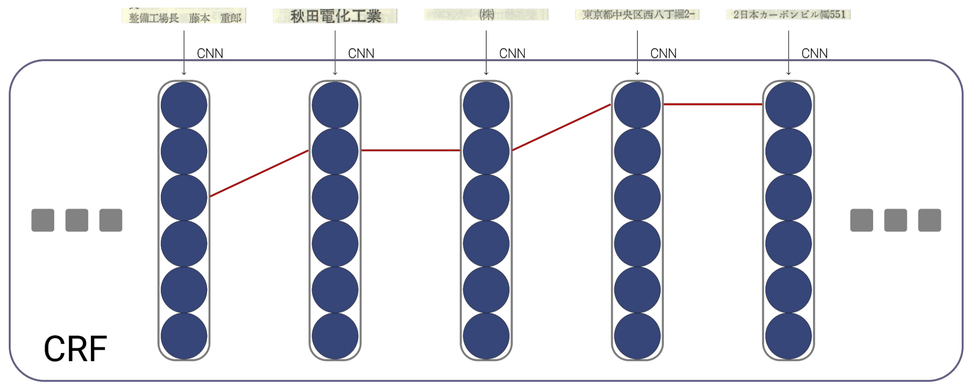
Segmentation

The purpose of segmentation is to 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).



Classification



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.

Experiment

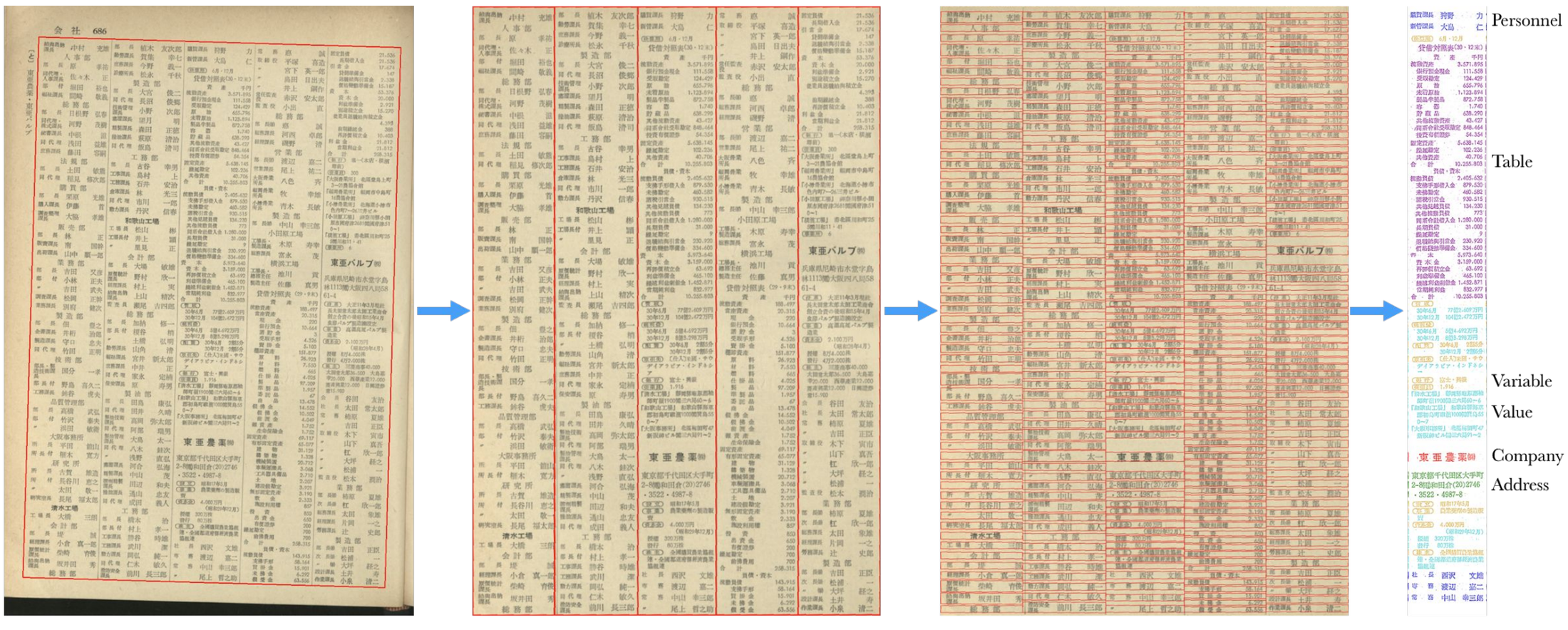
Data We randomly choose four text regions from PR1956 and manually label their row images (1407 row images in total) with correspondent class. Three text regions are used as the training (85%) and validation (15%) set, the other one text region is used as a test set. For the training and validation set, 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.

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

Method	CNN	CNN+CRF
Accuracy	95.6%	*96.8%*

Pipeline



scanned image with detected text region → column segmentation → row segmentation → row classification