Information Extraction from Text Regions with Complex Tabular Structure

Kaixuan Zhang *
Institute for Quantitative Social Science
Harvard University
Cambridge, MA 02138

Zejiang Shen
Institute for Quantitative Social Science
Harvard University
Cambridge, MA 02138

Jie Zhou
Department of Economics
MIT
Cambridge, MA 02139

Melissa Dell
Department of Economics
Harvard University
Cambridge, MA 02138

Abstract

Recent innovations have improved layout analysis of document images, significantly improving our ability to identify text and non-text regions. However, extracting information from within text regions remains quite challenging because the text region may have a complex structure. In this paper, we present a new dataset with complex tabular structure, and propose new methods to robustly retrieve information from the complex text region.

1 Introduction

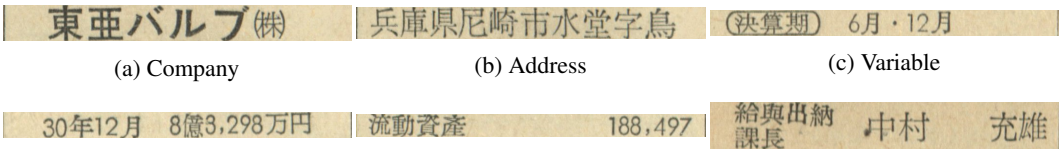
There has been an increasing interest in document analysis in recent years [1], with significant progress in automatic layout segmentation [3, 4] of document images into text and non-text regions. However, efforts to parse the information in text region have been limited. The challenge of parsing text 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This challenge is particularly severe in historical documents, as in the past document formats were much less standardized than today.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. For example, the Opportunity Insights initiative at Harvard cites a 29-million-dollar cost estimate to hand digitize historical census manuscripts that could uncover long-run patterns of economic mobility in U.S. communities. Because there has been limited progress in automating the digitization of historical tabular data, nearly all research on long-run economic phenomena uses aggregate data. Without historical data on individuals and firms, it is impossible to answer many fundamental questions, including those related to inequality, social mobility, and the role of firm-level factors in promoting economic innovation and growth. Automation has the potential to massively scale up the extraction of historical quantitative data, significantly expanding and democratizing access.

In this paper, we first present a dataset where the text region of document images has a complex tabular structure, then we propose a method for robustly retrieving information from this complex tabular structure, and finally we talk about future work. Our contribution is two-fold: firstly, we identify the challenges in parsing the text region and release a dataset to study this problem; secondly,

*kaixuanzhang@fas.harvard.edu

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Company (b) Address (c) Variable

(d) Value (e) Table (f) Personnel

Figure 1: Row category in PR1956.

Table 1: Category name and description.

Category Name	Description
Company	The name of a Japanese company recorded in PR1956.
Address	Address of correspondent company.
Variable	Information of correspondent company, such as total income, share holder, etc.
Value	Information of correspondent variable.
Table	One row in a table which reflects company asset.
Personal	A top manager and his position of correspondent company.

we propose a method to robustly parse the complex tabular structure of the text region and retrieve information.

2 Dataset

We collect around 1,000 scanned images of the 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 1956. We have applied image processing techniques on all scanned images to detect the text region, as well as segment columns and rows (Figure 2). Each text region have 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. However, if we apply google cloud vision (GCV), a state-of-art OCR API, to the text region directly, the text region is not parsed correctly since GCV fails to understand its complex tabular structure.

2.1 Row Category

In PR1956, we classify each row into six classes (Figure 1) depend on the content. Category names and their descriptions are listed in Table 1

3 Approach and Results

The pipeline of our method is shown in Figure 2. In order to parse the text region, we need to first segment the text region into columns and rows, and then classify row images into correspondent classes. Row segmentation and row classification give the structure of the text region, and finally we can simply apply OCR software on row images to get the text.

For column and row segmentation, we utilize basic image processing techniques (binarization, CCL, etc), and Figure 1 and 2 show some results of row segmentation. Due to the page limit we cannot discuss the techniques in detail, but more demos and our code is publicly available on Github [2].

Row classification is vitally important for extracting structured information from the text region. Moreover, it is extremely helpful for parsing row images of personnel where the left side of image may have multiple lines (e.g. Figure 1f). In the following section, we focus on how to robustly classify row images.

[2]https://github.com/KaixuanZ/PR1956

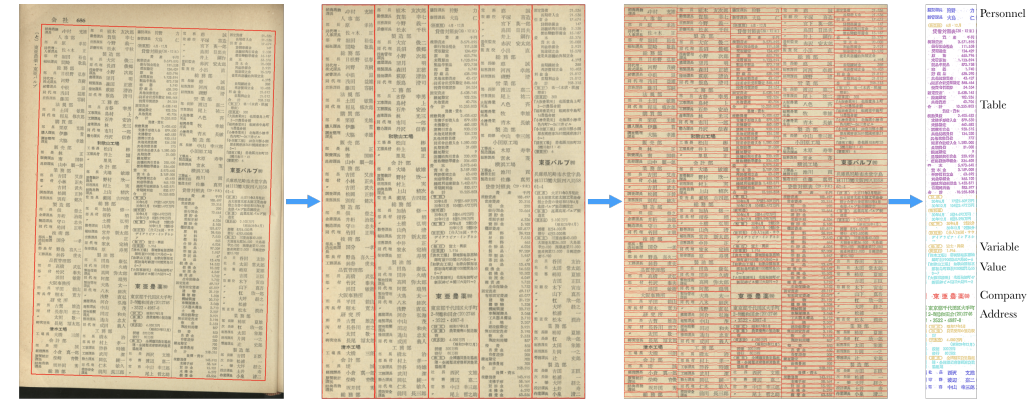


Figure 2: Pipeline of our method.

3.1 CNN for Row Classification

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 Model A MobileNet [2] is trained with our training data for 40 epochs. Input images are resized to 80*800. The neural net is initialized with parameters pre-trained by row images from another Japanese economic document, and then optimized by RMSProp [6]. The model with the highest validation accuracy is selected as the model for testing.

3.2 Linear-chain CRF for Sequence Analysis

There are rich information relationships between adjacent rows. For instance, addresses only appear after company names. We can view the row classification results as a sequence and improve the classification accuracy by applying a linear-chain conditional random field [5].

Emission Scores $U(\mathbf{x}_k, y_k)$ represents how likely is y_k given row image $\mathbf{x}_k$, and it is given by the trained CNN in section 3.1.

Transition Scores $T(y_k, y_{k+1})$ represents how likely is y_k followed by y_{k+1}. We first learn $T(y_k, y_{k+1})$ from our labeled data but no increase of accuracy is observed on our test data of PR1956, which is possibly because the training data (three text regions) is not representative enough to learn the sequence patterns. Therefore, we manually setup the transition scores by following equation:

$$T(y_k, y_{k+1}) = \begin{cases} 1 & if\ y_k\ can\ be\ followed\ by\ y_{k+1} \\ 0 & otherwise \end{cases}. \tag{1}$$

With these transition scores, we eliminate row pairs which cannot appear in PR1956. For instance, address can never be followed by company name. Finally, the CRF can be decoded by the Viterbi algorithm [7] to select the optimal sequence $\mathbf{y}$.

3.3 Results

Row classification accuracy is shown in Table 2. The CNN can achieve 95.6% accuracy because row images from different categories look quite different. For instance, company names (Figure 1a) are in bold font and table (Figure 1e) has a large blank space in the middle. As we expect, a higher classification accuracy is achieved when we combine the CNN and CRF, since the information between adjacent rows is considered in CRF.

Table 2: Classification accuracy on testset of PR1956

Method	CNN	CNN + CRF
Accuracy	95.6%	96.8%

4 Conclusion and Future Work

Our work provides a method to retrieve information from text regions with complex structures, and we explain how to classify row images in detail. In the future, we will investigate the following aspects:

Learning the Structure of Text Region For PR1956, we manually designed algorithms for column and row segmentation to derive the structure of the text region. It is worthwhile to design an algorithm that can automatically learn the structure.

Classification with Semantic Information In our current method, row image classification only utilizes information from the image. We are in the process of exploiting the semantic information from the OCR output in order to build a more detailed classifier.

References

[1] C. Clausner, A. Antonacopoulos, and S. Pletschacher. Icdar2017 competition on recognition of documents with complex layouts - rdcl2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1404–1410, Nov 2017.

[2] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[3] Yixin Li, Yajun Zou, and Jinwen Ma. Deeplayout: A semantic segmentation approach to page layout analysis. In *International Conference on Intelligent Computing*, pages 266–277. Springer, 2018.

[4] Sofia Ares Oliveira, Benoit Seguin, and Frédéric Kaplan. dhsegment: A generic deep-learning approach for document segmentation. *CoRR*, abs/1804.10371, 2018.

[5] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373, April 2012.

[6] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[7] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.