CORD: A Consolidated Receipt Dataset for Post-OCR Parsing

Seunghyun Park[1] **Seung Shin** **Bado Lee**

Junyeop Lee **Jaeheung Surh** **Minjoon Seo**

Hwalsuk Lee[2,*]

Clova AI, NAVER Corp.
Seongnam 13561, Korea
{seung.park[1], hwalsuk.lee[2]}@navercorp.com

Abstract

OCR is inevitably linked to NLP since its final output is in text. Advances in document intelligence are driving the need for a unified technology that integrates OCR with various NLP tasks, especially semantic parsing. Since OCR and semantic parsing have been studied as separate tasks so far, the datasets for each task on their own are rich, while those for the integrated post-OCR parsing tasks are relatively insufficient. In this study, we publish a consolidated dataset for receipt parsing as the first step towards post-OCR parsing tasks. The dataset consists of thousands of Indonesian receipts, which contains images and box/text annotations for OCR, and multi-level semantic labels for parsing. The proposed dataset can be used to address various OCR and parsing tasks.

1 Introduction

Optical character recognition (OCR) is a technique for converting images of characters into digitized texts [1, 2]. Recently, deep learning in computer vision domain has significantly improved the performances of OCR [3, 4]. Nonetheless, there is still huge room for improvement, especially concerning the tasks simultaneously linked to natural language processing (NLP) as well.

In particular, post-OCR parsing is currently one of the most important, yet challenging problems in both OCR and NLP community. The goal of post-OCR parsing is to predict pre-defined semantic labels from the given OCR. Researchers from both domains have long tried to tackle the problem and collected a significant amount of data sets independently. However, since it is a specialized task, the datasets contain critical limitations to provide proper supervision. The OCR datasets typically do not have parsing class labels for the extracted texts. The parsing datasets usually contain error-free and well-ordered digitized texts in contrast to the erroneous outcomes from OCR process. We can add synthetic noise to the parsing data, but the distribution and error patterns could be different from the OCR errors, which would inevitably lead to the degradation of generalization performance.

Over the past few years, a few post-OCR parsing datasets have been made public through post OCR challenges [5]. For example, ICDAR 2019 Post-OCR Challenge introduced the Scanned Receipts OCR and Information Extraction (SROIE) dataset [6]. It provides receipt images of texts and two types of annotations for OCR and parsing problem: (1) box-level text annotations for OCR, and (2)

*To whom correspondence should be addressed.

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

document-level parse annotations for parsing. Although the availability of both OCR and parsing information have given rise to active research within the field, it still possesses some shortcomings, *e.g.*, limited data size and lack of box-level parsing annotations. Considering that only hundreds of samples are provided in the SROIE dataset, weak document-level annotations could not provide enough supervision for training a model with satisfactory performance.

In this paper, we introduce a novel dataset called **CORD**, which stands for a **Co**nsolidated **R**eceipt **D**ataset for post-OCR parsing. To the best of our knowledge, this is the first publicly available dataset which includes both box-level text and parsing class annotations. The parsing class labels are provided in two-levels. The eight superclasses include *store*, *payment*, *menu*, *subtotal*, and *total*. The eight superclasses are subdivided into 54 subclasses *e.g.*, *store* has nine subclasses including *name*, *address*, *telephone*, and *fax*.

Furthermore, it also provides line annotations for the serialization task which is a newly emerging problem as a combination of the two tasks. Current semantic parsing techniques can handle only well-ordered texts. Texts obtained by OCR, however, are in two-dimensional space, thus we need an appropriate serialization technique for mapping obtained texts into one-dimensional space. In our experiments, serialization has a significant impact on parsing performance.

To recapitulate briefly, the key contributions of our paper are as follows:

- We introduce a novel and large-scale receipt dataset that can be used for OCR and parsing tasks, from task-specific to end-to-end.
- Our dataset provides multi-level labels for weakly and strongly supervised parsing tasks.

The dataset and descriptions will be available on `https://github.com/clovaai/cord` at the time of publication.

2 Data Acquisition

2.1 Annotation

We collected over 11,000 Indonesian receipt images through crowd-sourcing. Receipts have been obtained from shops and restaurants. In general, crowd-sourcing involves providing a guideline and annotation tool for crowd workers. For making guidelines, we first sampled hundreds of receipts and analyzed their common structures. The sampled receipts are carefully examined and then dominant and useful parse categories were defined. The main parse classes consist of store information, payment information, menu, void menu, subtotal, void total, total. Since these structures are not easily perceptible for the human mind, we carried out pilot annotation and redesigned parse categories repeatedly. Finally, we created two guidelines for the OCR and parse annotation.

For more efficient annotation, we developed a web-based annotation tool and provided it to crowd workers. The annotation tool has two types of user accounts: annotator and inspector. When the annotator completes the annotation on an image, the inspector checks the annotation results and determines whether the annotator complies with the guidelines or not. If the annotation is not done properly, the corresponding image is reassigned to another annotator.

2.2 Anonymization

Although a relatively small amount, receipts may have sensitive information, *e.g.*, customer's name, number of credit/debit card, transaction date/time. We thoroughly inspected the image and then removed the information by blurring in the image and deleting the corresponding field in the JSON-formatted file.

3 Data Specification

3.1 Structure of Data

The receipts dataset consists of more than 11,000 image and JSON pairs. An example of image and json pair is shown in Figure 1. The ground truth has three main attributes, `meta`, `the region of`

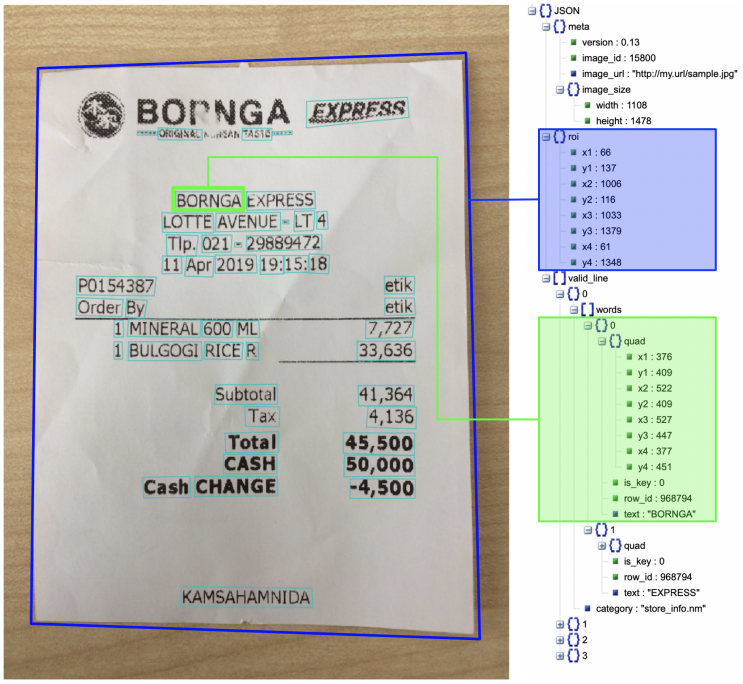

Figure 1: An example of receipt image (left) and json (right).

Table 1: Dataset statistics

No.	Superclass	# Subclasses	Proportion	Example
1	store info.	9	0.134	store name, address, telephone number
2	payment info.	2	0.092	visiting time, card company
3	menu	16	0.510	menu name, quantity, price, submenu
4	void menu	6	0.0002	menu name, quantity, price
5	subtotal	8	0.073	subtotal price, discount, service charge, tax
6	total	8	0.145	total price, amount of credit/debit card
7	void total	4	0.00015	void total, void tax
8	etc.	1	0.045	table number, membership points, repeated symbols
	Total	54	1.0	

`interest` (`ROI`), and `valid line`. The `meta` field holds the overall information of the image, such as `image size`, `image url`, and `image id`. (`ROI`) contains four coordinates that encompass the area of the receipt.

The `valid line` field has crucial information for post-OCR parsing. The `quad` field contains four coordinates of quadrilateral, and the `text` field has the incorporating text of the corresponding box. `quad` and `text` fields are used for OCR detection/localization and recognition task, respectively. Note that only optically identifiable text instances are annotated.

For parsing tasks, there are three additional fields, `category`, `is_key`, and `row_id`. The `category` indicates parse class label. Note that the details of parse classes are explained on the Section 3.2. The `row_id` is an index of the line. The text instances which have the same `row_id` are on the same line. As represented in Figure 1, *BORNGA* and *EXPRESS* have the same `row_id` since they are placed next to each other. The `is_key` flag is used to identify words that act as a key to other text elements. For example, the `is_key` value of the text *BORNGA* is 0 since it does not act as a key but a value. On the other hand, the text *Total* has an `is_key` value of 1 because it acts as a key to another text element, *45,500*.

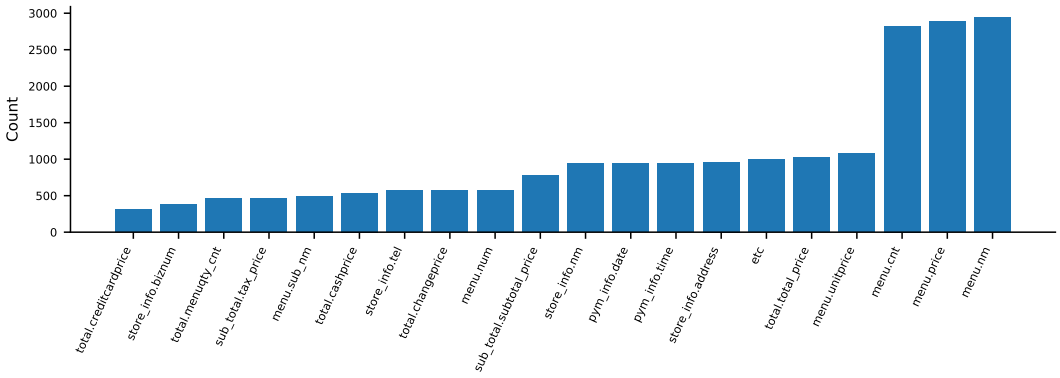

Figure 2: The top 20 class labels within randomly sampled 1,000 data examples

3.2 Definition of Parse Classes

We present dataset statistics in Table 1. A dataset consists of eight superclasses: *store*, *payment*, *menu*, *void menu*, *subtotal*, *void total*, *total*, and *etc*. These eight superclasses are divided into 54 subclasses, *e.g.*, the superclass *menu* has 16 subclasses containing *menu name*, *quantity*, *unit price*, *discount price*, *submenu*. Figure 2 represents the top 20 class labels from 1,000 randomly sampled data examples. The most frequent class label is the *menu*, especially *menu name*, *price*, and *count*.

4 Concluding Remarks

In this paper, we introduce a novel receipt dataset for a unified OCR-parsing task. The proposed dataset can be exploited not only each task-specific task but also end-to-end approaches. A receipt is a complex type of document. It contains many numbers and symbols as well as plain text, and it also has a complex text layout such as a tablet-shaped layout. Besides, receipt images acquired through the OCR process have various types of optical noise that originated from wrinkles and warpage. Due to these characteristics, the proposed data is more suitable for the OCR-parsing task than the synthetically generated dataset in terms of generalization performance. As mentioned in Section 3.2, our dataset provides multi-level class labels, for eight superclasses and 54 subclasses. It can be used to carry out weakly supervised learning as well as strongly supervised learning.

Note that we will first release 1,000 samples which consist of 800 (train), 100 (dev), and 100 (test) data examples, and the remaining data will be published in sequence. The exposure of sensitive information can cause a legal problem, so we will carefully examine and remove sensitive information.

Acknowledgments

Special thanks to Line Corporation, Indonesia and the members of Clova AI OCR.

References

[1] Zhou, X. et al. East: an efficient and accurate scene text detector. In *CVPR*, pages 2642–2651, 2017.

[2] Liao, M., Shi, B. and Bai, X. Textboxes++: A single-shot oriented scene text detector. *Image Processing*, 27 (8):3676–3690, 2018.

[3] Baek, J. et al. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[4] Baek, Y. et al. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[5] Karatzas, D. et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[6] Karatzas, D. et al. The robust reading competition annotation and evaluation platform. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 61–66. IEEE, 2018.