Supplementary Materials of D^{4}LA Dataset

Cheng Da*, Chuwei Luo*, Qi Zheng, and Cong Yao†
DAMO Academy, Alibaba Group, Beijing, China
dc.dacheng08,luochuwei,zhengqisjtu,yaocong2010@gmail.com

A. Annotations of D^{4}LA Dataset

It is time-consuming and labor-intensive to manually annotate the images of various document types with complex layout categories. We employ about 5 full-time annotators to annotate these complex document images for about 1.5 months. The definition of categories and the guideline of annotations are carefully designed and can be basically applied to other types of documents. The layout annotations of the bounding boxes in our D^{4}LA dataset are in standard MSCOCO format for the classic detection task.

For the OCR results of document images in our D^{4}LA dataset, we first map the images of D^{4}LA into RVL-CDIP dataset, and further, map them into IIT-CDIP dataset which is the superset of RVL-CDIP and provides the text contents and bounding boxes for each word. The original images, the OCR results of them and the manual layout annotations will be made publicly available.

B. Detailed Layout Categories in D^{4}LA

We describe the definitions of the different layout categories of the proposed D^{4}LA dataset. We simply introduce the common categories of scientific papers which is similar to those of DocBank. For some special layout categories of D^{4}LA, we illustrate them in detail.

B.1. Common Categories in Scientific Papers

Documents of the existing large-scale DLA datasets are mainly scientific papers. The layout categories are defined especially for scientific papers. These common categories are not only included in D^{4}LA but also more detailed. We introduce these common categories as follows:

DocTitle is the title of the document that is similar to *Title* of papers in DocBank. While, in other types of documents, we define the text at the head of the document as "DocTitle", that is commonly bold or with underlines.

ListText is a paragraph with bullet or enumeration symbols, which is different from *List* in PubLayNet and DocBank. Specifically, *List* is a region where all instances of "ListText" are grouped together into one "List" object block.

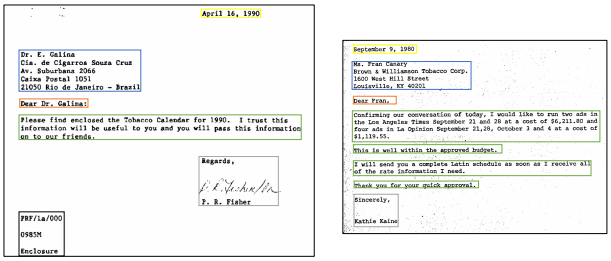

Figure 1. Some special layout categories of letters in D^{4}LA. Best viewed in color.

While "ListText" is an individual object instance that is one of the paragraphs of *List* region. The definition of "ListText" is more suitable for other types of documents, since "ListText" instances are often mixed with other text.

Table and Figure are common object instances in documents as in PubLayNet and DocBank.

TableName and FigureName are the captions of the Table and Figure, respectively. While they are both *Caption* in DocBank.

Footer is the footnote of the document, which often begins with special symbols.

PageHeader and PageFooter are the page header and page footer on the page, respectively.

Author represents the author of the paper or other documents, *e.g.* , News article, Scientific report.

Abstract often appears at the beginning of the paper behind a section of "Abstract" or "Summary".

ParaText is a paragraph that may have multiple lines when the paragraph is long. Notably, "ParaText" is different from "ListText" which contains special enumeration symbols.

ParaTitle is similar to *Section* in DocBank, which is the title of one paragraph of "ParaText".

Equation is the formula or equation in the paper, that often includes formula numbers.

Reference often includes a reference number, authors, article name, journal name, page number, dates, and so on. All references constitute a "Reference" region block.

PageNumber is the page number of a document that often appears in the header or footer of the page.

OtherText represents some text with word phrases that is

*Equal contribution. † Corresponding author.

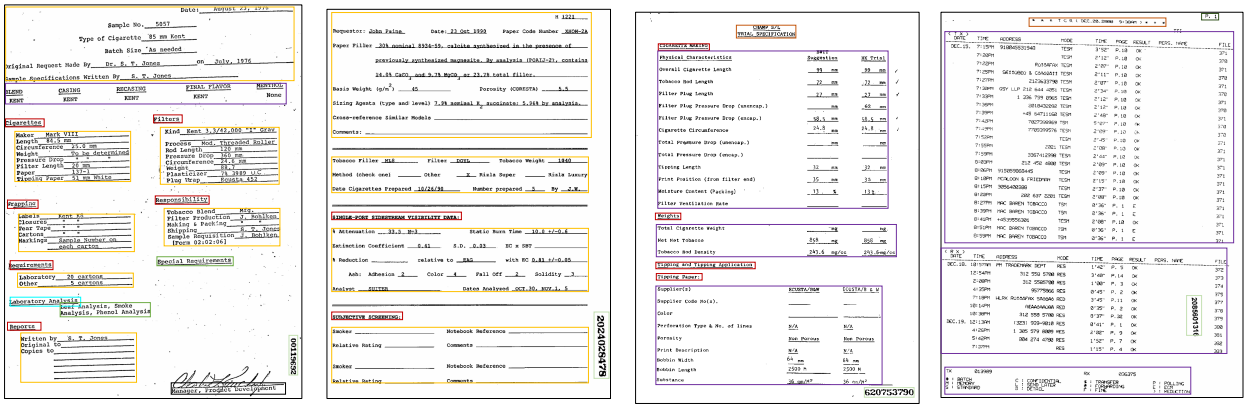

Figure 2. Some special layout categories of forms in D^{4}LA. Best viewed in color.

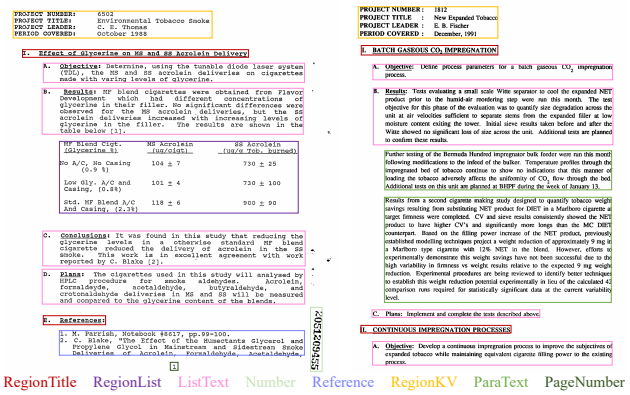

Figure 3. Comparison between different types of text.

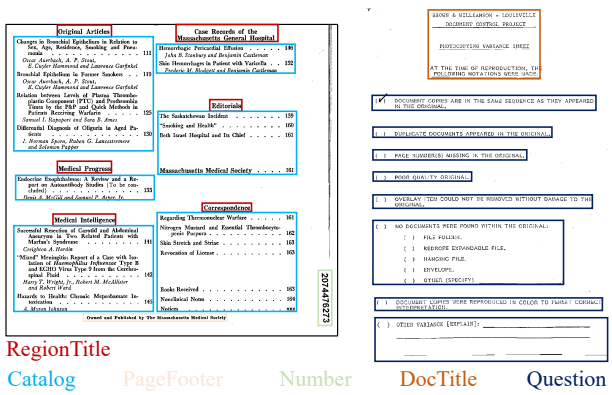

Figure 4. Other special layout categories. Best viewed in color.

not a complete paragraph and is not belong to any other layout categories. *e.g.* , some useless text.

B.2. New Categories in Letters

By analyzing the documents of letters in RVL-CDIP, we observe that a standard letter usually has a fixed format. We customize 3 classes, *i.e.* , LetterHead, LetterDear and LetterSign for documents of letters, as shown in Figure 1.

LetterHead represents the inside address that often appears at the beginning of the letter and records the name and address of the recipient.

LetterDear is the salutation or greetings to the recipient, which is usually behind the "LetterHead".

LetterSign includes the complimentary close and signature, which is often at the end of letters.

Date often appears in letters and papers that include years, months, and days.

B.3. New Categories in Forms

Scientific publications are mostly composed of regular paragraphs, tables and figures. While other documents often include irregular areas, such as the Key-Value pairs in invoices or the line-less list areas in budget sheets. This semi-structured data is more important than ordinary words in the document for downstream works, such as information extraction. Thus, we define 3 region blocks for special use. Some cases are illustrated in Figure 2.

RegionKV is a region that contains Key-Value areas.

RegionList is a region that includes wireless form or lineless list areas.

RegionTitle is the title of the complex region, *e.g.* , "RegionKV", "RegionList" and "ListText", which is different from "ParaTitle" of a paragraph. Typically, both "ParaTitle" and "RegionTitle" may contain enumeration symbols, which may be confused with "ListText". Thus, distinguishing between these texts requires incorporating the semantics of the context. We show two difficult cases in Figure 3.

B.4. Other Categories

The other remaining categories are shown in Figure 4.

Number represents the special number in IIT-CDIP that is not the content of the document and often vertical text.

Catalog includes text and page numbers, which is a region block not one text line with the page number.

Question often appears in the questionnaire. They are mostly true or false questions in D^{4}LA dataset.