Information Extraction - Structured Data

Introduction

The goal of an information extraction system is to find and understand limited relevant parts of the information where the system gathers information from many pieces of text. The goal of doing this is to produce a structured representation of relevant information. The documents, such as forms, receipts, bills, insurance quotes, and others, are extremely common and critical in a diverse range of business workflows. Processing these documents manually requires a good amount of effort. Though automated document processing methodologies exist, they require more research improvement.
Consider a document type like invoices, which can be laid out in thousands of different ways — invoices from different companies, or even different departments within the same company, may have slightly different formatting. However, there is a common understanding of the structured information that an invoice should contain, such as an invoice number, an invoice date, the amount due, the pay-by date, and the list of items for which the invoice was sent. A system that can automatically extract all this data has the potential to dramatically improve the efficiency of many business workflows by avoiding error-prone, manual work.

Dataset and Annotation

Refer here for the dataset.

The dataset has 58 ‘’990 form” sample pdfs with different document types. Each form contains different fields like those given below in json format.

```
{  "data": {
"Basic Information": {
        "Tax_Beginning_Date": {
        "value": "01/01/2015",
        "position": [
        634.2,
```

287.9,
780.5,
314.5
],
"confidence": 1.0,
"review_required": false
},
"Tax_Ending_Date": {
"value": "31/12/2015",
"position": [
944.6,
274.6,
1095.4,
323.4
],
"confidence": 1.0,
"review_required": false
},
"Address_Change": {
"value": false,
"position": [
4.4,
361,
44.3,
398.7
],
"confidence": 1.0,
"review_required": false
},
"Name_Change": {
"value": false,
"position": [
11.1,
405.3,
46.6,
440.7

```
],
"confidence": 1.0,
"review_required": false
},
"Initial_Return": {
"value": false,
"position": [
6.7,
454,
48.8,
485
],
"confidence": 1.0,
"review_required": false
},
"Final_Return/Terminated": {
"value": false,
"position": [
8.9,
509.4,
46.6,
551.5
],
"confidence": 1.0,
"review_required": false
},
"Amended_Return": {
"value": false,
"position": [
6.7,
560.3,
48.8,
602.4
],
"confidence": 1.0,
"review_required": false
```

```
},
"Application_Pending": {
"value": false,
"position": [
13.3,
609.1,
39.9,
633.4
],
"confidence": 1.0,
"review_required": false
},
"Organization_Name": {
"value": "WINNINGHABITS CHARITABLE FOUNDATION",
"position": [
250.6,
347.7,
707.3,
376.5
],
"confidence": 1.0,
"review_required": false
},
"Organization_Street_Address": {
"value": "4755 CHAPEL HILL",
"position": [
252.8,
513.8,
501.1,
544.8
],
"confidence": 1.0,
"review_required": false
},
"Organizarion_City/State/Country/Zip": {
"value": "DALLAS, TX 75214",
"position": [
257.2,
589.1,
465.6,
626.8
],
"confidence": 1.0,
"review_required": false
},
"Employer_Identification_Number": {
"value": "01-0887302",
"position": [
1246.1,
361,
1432.4,
414.2
],
"confidence": 1.0,
"review_required": false
},
"Phone_Number": {
"value": "(214) 363-6586",
"position": [
1250.6,
518.3,
1470.1,
553.7
],
"confidence": 1.0,
"review_required": false
},
"Website": {
"value": "WWW FAMILY BUILD ORG",
"position": [
195.1,
863.8,
534.4,
890.3
],
"confidence": 1.0,
"review_required": false
},
"Corporation_Type_Organization": {
"value": true,
"position": [
241.7,
925.8,
277.2,
950.1
],
"confidence": 1.0,
"review_required": false
},
"Trust_Type_Organization": {
"value": false,
"position": [
392.5,
914.7,
432.4,
952.4
],
"confidence": 1.0,
"review_required": false
},
"Association_Type_Organization": {
"value": false,
"position": [
483.4,
921.3,
521.1,
954.6
],
```

"confidence": 1.0,
"review_required": false
},
"Other_Type_Organization": {
"value": false,
"position": [
627.5,
916.9,
663,
952.4
],
"confidence": 1.0,
"review_required": false
},
"Cash_Accounting_Method": {
"value": "",
"position": [],
"confidence": 1.0,
"review_required": false
},
"Accural_Accounting_Method": {
"value": "",
"position": [],
"confidence": 1.0,
"review_required": false
},
"Other_Accounting_Method": {
"value": "",
"position": [],
"confidence": 1.0,
"review_required": false
},
"Tax-ExemptStatus501(c)(3)": {
"value": true,
"position": [
257.2,

```
804,
303.8,
854.9
],
"confidence": 1.0,
"review_required": false
},
"Tax-ExemptStatus501(c)(19)(insertno)": {
"value": false,
"position": [
410.2,
812.8,
441.3,
848.3
],
"confidence": 1.0,
"review_required": false
},
"Tax-ExemptStatus4947(a)(1)": {
"value": false,
"position": [
725.1,
815,
765,
846
],

},
"Tax-ExemptStatus527": {
"value": false,
"position": [
911.3,
810.6,
946.8,
848.3
],
                }
        }
        }
```

Task

Key Information Extraction from 990 Forms: The aim of this task is to extract texts of a number of key fields from given forms, and save the texts for each form pdf in a json file.

Evaluation Protocol

For each form sample, the extracted text is compared to the ground truth. An extract text is marked as correct if both submitted content and category of the extracted text match the ground truth; Otherwise, marked as incorrect. The precision is computed over all the extracted texts of the test form samples. F1 score is calculated based on precision and recall. F1 score will be used for ranking.