Visual Context and Natural Language
An ideal text processing pipeline for better accuracy and document modeling

Gowri Sekar
Agenda

- Documents And Their Flavors
- An Ideal Text Extraction Pipeline
- Visual Context From Raw Documents
- Natural Language Models Upon Visual Context
- Enhanced Operations with the Model
- The stack for a digitization warehouse
- Questions
Documents and Their Flavors

So what is the huge deal about them?
Gowri Investments

Sekar
653 Humming Avenue, Red Robin Township
Bangalore - 560001

STATEMENT

STATEMENT NUMBER | 12345663
STATEMENT DATE   | April 15, 2018
ACCOUNT NUMBER   | 654321456332
AMOUNT DUE       | 500.00
PAYMENT DUE DATE | 30-Apr-18

<table>
<thead>
<tr>
<th>DATE</th>
<th>DESCRIPTION</th>
<th>CHARGES</th>
<th>CREDIT</th>
<th>BALANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500.00</td>
</tr>
</tbody>
</table>

BALANCE FORWARD:

<table>
<thead>
<tr>
<th>DATE</th>
<th>DESCRIPTION</th>
<th>CHARGES</th>
<th>CREDIT</th>
<th>BALANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>600.00</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Please pay last amount in Balance column.

Please detach portion below and send with your remittance.

Sekar
653 Humming Avenue, Red Robin Township
Bangalore - 560001

ADDRESS QUESTIONS TO
Gowri Investments
Address Line1 Road
Address Line2 Road
Bangalore - 560000

MAKE ALL CHECKS PAYABLE TO
Gowri Investments

EMAIL QUERIES TO
gowriinv@ginv.com
RENTAL AGREEMENT

This agreement of rent is made and executed at Bangalore on this 1st day of April, 2009 (01-04-2009) by and between:-

Mr. Freeze
189 West Avenue,
Hot Cross Street, Mumbai - 87

Hereinafter called the LESSOR/OWNER of the one part; and in favor of:-

Mr. P. Scarecrow
Working at:
Harley Corporation
789 Officers Lane
Mumbai -68

Permanent Address:
D. No: 9-109
Hatter 3rd Phase
Bank Lane
BANGALORE - 71

Hereinafter called the LESSEE/TENANT of the other part; witnesses as follows:-

Whereas the term lessor and lessee shall mean and include their respective heirs, legal representatives, administrators, and assigns, etc.

And whereas the lessor is the sole and absolute owner of the premises situated at 189 West Avenue, Hot Cross Street, Mumbai – 87 and where as the lessee has approached with the lessor to let-out the schedule premises on rental basis. And the lessor agrees to let-out the same under the following terms and conditions:-

1. whereas the lessor agrees to let-out the schedule premises for a monthly rent of 7000/- (Rupees Seven Thousand only) per month, and the lessee shall agrees to pay the same on every month.

2. Whereas the lessee shall agrees to pay the monthly rent by on or before 5th of Every English calendar month.

3. Whereas the lessee shall use the schedule premises for residential purpose only and should not use for any illegal or immoral purposes.
An Ideal Text Extraction Pipeline

Bring it On!
An Ideal Text Extraction Pipeline

[Diagram showing steps of text extraction pipeline]

1. Bounding Boxes
2. Vertical Groups
3. Sections
4. Tables
5. Paragraphs

[Visual Context Model]

- TEM Grammar
- NLP Model

[Boxes containing:
- Basic NER
- Address Detection
- Currencies and Money
- Domain Specific Knowledge Base]
Visual Context – Key Principles

**Visual Bounding Boxes**

Our eyes are generally trained to look for boundary lines to mark end and start of objects. Let's use that concept.

- Find alignment boundary lines in the document by finding elements which align.
- For each rectangular element in document:
  - If it is surrounded by actual lines, use those to create its boundary box.
  - Else use alignment boundary lines if present
  - Else fall back to mid-point from the nearest neighbor for boundaries

**Establishing Traversal**

Methods to answer queries like – provide all the elements below an element, left / right of an element etc.
# Gowri Investments

**STATEMENT**

| ADDRESS QUESTIONS TO | Gowri Investments  
|----------------------|------------------ |
|                     | Address Line1 Road  
|                     | Address Line2 Road  
|                     | Bangalore - 560000 |

| STATEMENT NUMBER | 12345663  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STATEMENT DATE</td>
<td>April 15, 2018</td>
</tr>
</tbody>
</table>
| ACCOUNT NUMBER   | 654321456332  
| AMOUNT DUE       | 500.00     |
| PAYMENT DUE DATE | 30-Apr-18   |

<table>
<thead>
<tr>
<th>DATE</th>
<th>DESCRIPTION</th>
<th>CHARGES</th>
<th>CREDIT</th>
<th>BALANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500.00</td>
</tr>
</tbody>
</table>

**Please pay last amount in Balance column.**

---

**STATEMENT NUMBER** | 12345663  
**PAYMENT DUE DATE** | April 30, 2018 |
**ACCOUNT NUMBER**   | 654321456332  
**AMOUNT DUE**       | 500.00     |
**AMOUNT ENCLOSED**  |            |

**MAKE ALL CHECKS PAYABLE TO** | Gowri Investments  
**EMAIL QUERIES TO** | gowriinv@ginv.com
RENTAL AGREEMENT

This agreement of rent is made and executed at Bangalore on this 1st day of April, 2009 (01-04-2009) by and between:-

Mr. Freeze
189 West Avenue,
Hot Cross Street, Mumbai - 87

Hereinafter called the LESSOR/OWNER of the one part; and in favor of:-

Mr. P. Scarecrow
Working at:
Harley Corporation
789 Officers Lane
Mumbai - 68

Hereinafter called the LESSEE/TENANT of the other part; witnesses as follows:-

Whereas the term lessor and lessee shall mean and include their respective heirs, legal representatives, administrators, and assigns, etc.

And whereas the lessor is the sole and absolute owner of the premises situated at 189 West Avenue, Hot Cross Street, Mumbai - 87 and where as the lessee has approached with the lessor to let-out the schedule premises on rental basis. And the lessor agrees to let-out the same under the following terms and conditions:-

1. Whereas the lessor agrees to let-out the schedule premises for a monthly rent of 7000/- (Rupees Seven Thousand only) per month, and the lessee shall agrees to pay the same on every month.

2. Whereas the lessee shall agrees to pay the monthly rent by on or before 5th or Every English calendar month.

3. Whereas the lessee shall use the schedule premises for residential purpose only and should not use for any illegal or immoral purposes.
The Model as Rules

Rules Grammar:

- On 1st page
- Top-right quadrant
- Around keyword "Invoice #" or similar
  - {Tax Invoice No}
- Right of keyword, within 1 to 9 words
  - {1 word}
- With 4+ characters
  - {15 characters}
- Contains numbers or special characters
  - {numbers}
The Model as an Enabler

- Extraction with Improved Accuracy
- Search Formats
- Heuristics on Layouts
- Fraudulent Documents
- Probabilistic Models within Visual Context
- Heuristics on Content

DIGITIZATION MODEL
Text Extraction Component Stack

API Layer (Read, Write, Extract, Train, Search)

- Text Extraction Engine
- Training Module
- Exception Engine
- Confidence Refiner

Rules Model

Training Dataset

Versioned Storage – Document and Metadata

Indexed/Projected Storage

Metadata Model
Questions?
Disclaimer:
Views and opinions expressed are for informational purposes only. They do not constitute a recommendation by Goldman Sachs to buy, sell or hold any security. No part of this material may, without Goldman Sachs’s prior written consent, be (i) copied, photocopied or duplicated in any form, by any means, or (ii) distributed to any person that is an employee, officer, director, or authorized agent of the recipient.

Copyright Note: 2018 Goldman Sachs. All rights reserved.
GREAT INDIAN DEVELOPER SUMMIT 2019
Conference: April 23-26, Bangalore

Register early and get the best discounts!

www.developersummit.com @greatindiandev bit.ly/gidslinkedin