# INFORMATION EXTRACTION -THE EASY WAY

# OUTCOMES FROM THIS TALK

**THE PROBLEM**

To extract required information from scanned pdf documents easily

Why this problem is complex

Approaches to solve this problem

How we are going to get a solution to this problem in a easier way

How efficient is our solution

Who can find this solution useful

**UNDERLYING TECH STUFF**

What is OCR

Identifying the required data from document

Extracting as key,value pairs

Underlying algorithms and logic

How to install and run

How to use this tool

Enhancements

# HELLO!

I am Sravya, working in Pramati Technologies Hyderabad

I am here to share my thoughts to solve the problem of extracting text from scanned pdfs.

You can find me at @sravya_ysk

# 1.

# INFORMATION EXTRACTION

Let's start with it in a easier way !!

"DEALING WITH DOCUMENTS AND EXTRACTING TEXT OUT OF IT IS DEMANDING"

# EXISTING APPROACH

- Manual Data Entry
- OCR
- Rules or Template based extraction

The flipside of this approach is if document deviates from expected template the defined rules would not work.

# COMPLEXITIES INVOLVED

- ❑ Variance in format of documents
- ❑ Quality of documents
- ❑ Hand written text, scribblings in document
- ❑ Variance in the fields that you are interested to extract
- ❑ Quality of output of OCR

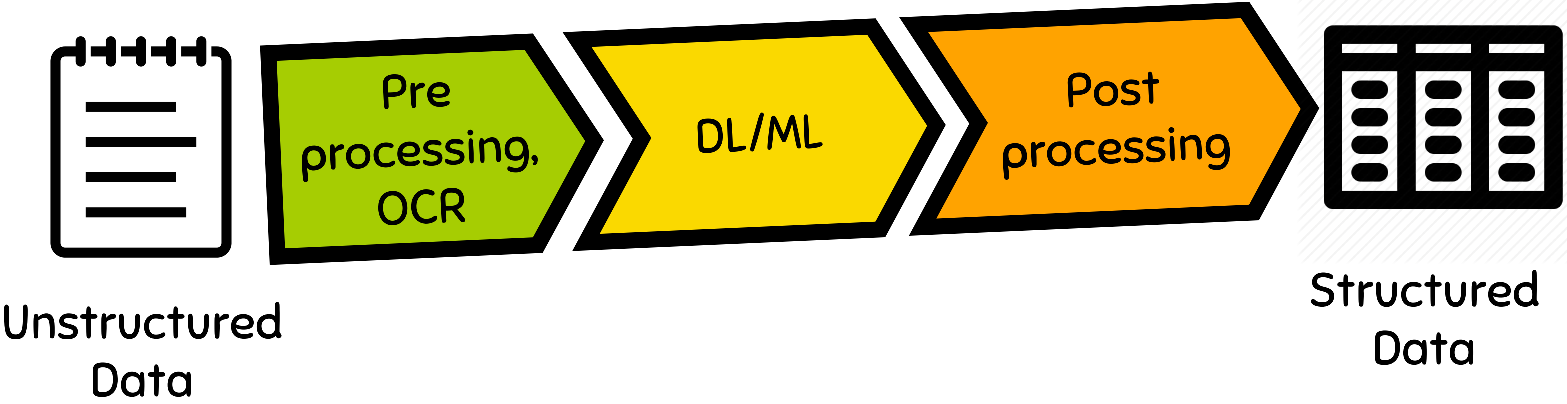
# OUR SOLUTION

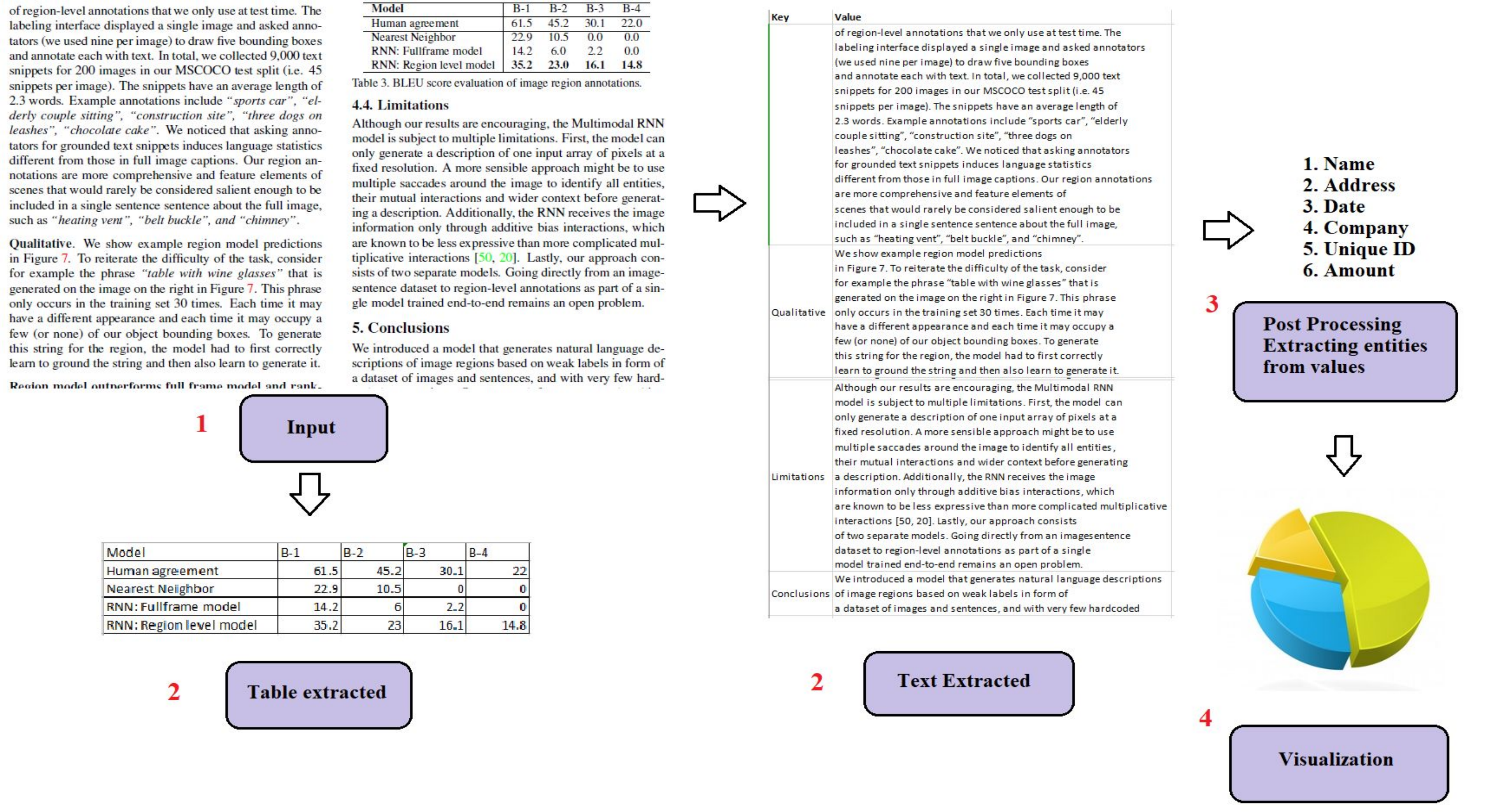
Automatic extraction of text/data from documents like scanned pdfs, word/excel files without users manual effort

# PIPELINE

- Take Raw Input Images
- File segregation
- Image Annotations (Bounding boxes) **in training phase**
- Text Detection – Text Localisation / Document Orientation Analysis
- Text Recognition – OCR
    a. Tesseract
    b. Abbyy
- Text Structuring
- Text-Cleaner/ Binarization
- Classification (ML / Statistical Inference /DL)
- Table/Text extraction
- Post processing
- Entity extraction
- Data Store
- Visualization

# OUR PROCESS IS EASY

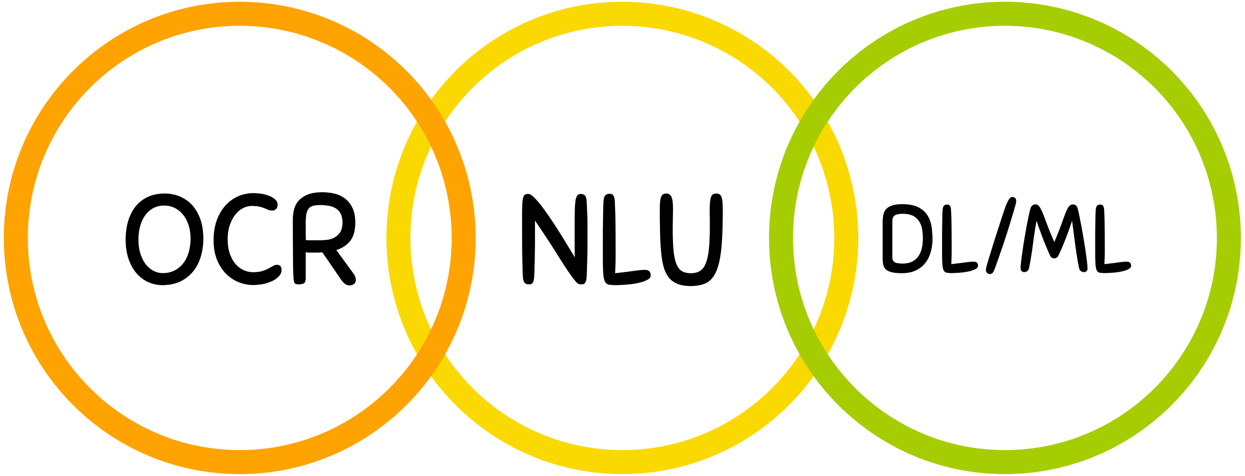



of region-level annotations that we only use at test time. The labeling interface displayed a single image and asked annotators (we used nine per image) to draw five bounding boxes and annotate each with text. In total, we collected 9,000 text snippets for 200 images in our MSCOCO test split (i.e. 45 snippets per image). The snippets have an average length of 2.3 words. Example annotations include "sports car", "elderly couple sitting", "construction site", "three dogs on leashes", "chocolate cake". We noticed that asking annotators for grounded text snippets induces language statistics different from those in full image captions. Our region annotations are more comprehensive and feature elements of scenes that would rarely be considered salient enough to be included in a single sentence sentence about the full image, such as "heating vent", "belt buckle", and "chimney".
Qualitative. We show example region model predictions in Figure 7. To reiterate the difficulty of the task, consider for example the phrase "table with wine glasses" that is generated on the image on the right in Figure 7. This phrase only occurs in the training set 30 times. Each time it may have a different appearance and each time it may occupy a few (or none) of our object bounding boxes. To generate this string for the region, the model had to first correctly learn to ground the string and then also learn to generate it.
Model B-1 B-2 B-3 B-4
Human agreement 61.5 45.2 30.1 22.0
Nearest Neighbor 22.9 10.5 0.0 0.0
RNN: Fullframe model 14.2 6.0 2.2 0.0
RNN: Region level model 35.2 23.0 16.1 14.8
Table 3. BLEU score evaluation of image region annotations.
4.4. Limitations
Although our results are encouraging, the Multimodal RNN model is subject to multiple limitations. First, the model can only generate a description of one input array of pixels at a fixed resolution. A more sensible approach might be to use multiple saccades around the image to identify all entities, their mutual interactions and wider context before generating a description. Additionally, the RNN receives the image information only through additive bias interactions, which are known to be less expressive than more complicated multiplicative interactions [50, 20]. Lastly, our approach consists of two separate models. Going directly from an image-sentence dataset to region-level annotations as part of a single model trained end-to-end remains an open problem.
5. Conclusions
We introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard-
1
Input
Model B-1 B-2 B-3 B-4
Human agreement 61.5 45.2 30.1 22
Nearest Neighbor 22.9 10.5 0 0
RNN: Fullframe model 14.2 6 2.2 0
RNN: Region level model 35.2 23 16.1 14.8
2
Table extracted
Key Value
Qualitative
Limitations
Conclusions
2
Text Extracted
1. Name
2. Address
3. Date
4. Company
5. Unique ID
6. Amount
3
Post Processing Extracting entities from values
4
Visualization

# FEATURES

1. Accepts scanned pdfs, word files, excel files as input
2. Efficient ensembled OCR engine
3. Automatic table extractor
4. Intelligent text extractor
5. Automatic document processing without writing specific rules
6. Extracts all identified text as key-value pairs
7. Entitiy extraction on top of extracted text
8. Currently we have data trained for Insurance, Health care, education, market slips etc

# SOME TECH STUFF BEHIND THIS TOOL



# COMPLEXITIES INVOLVED IN DEVELOPMENT

| | |
|---|---|
| Manual Preparation of Train Data | Handling large volumes of unstructured data |
| Train data related to some domains | Parallel processing |
| Context based text extraction | Manual testing |

# FUTURE ENHANCEMENTS

- ★ Enhancing document quality using GANs
- ★ Increasing the scope of training data to multiple domains
- ★ Enhancing OCR output
- ★ Increasing model efficiency in extraction task

# 1,26,124

No. of files processed across multiple domains

# 75%

Reduction in run time & memory efficient

# 90%

Accuracy !

# WOO HOO !

Now, you can extract required information from documents like contracts, tax documents, sales orders, bills, enrollment forms, benefit applications, insurance claims, policy documents, market slips, medical documents etc irrespective of the template of document

# THANKS!

Any questions?

You can find me at @sravya_ysk & sravyaysk@gmail.com

# CREDITS

Special thanks to all the people who made this talk possible

Do reach out to me @sravya_ysk

- Sravya.Y
- Pramati Technologies,Hyderabad