

Abstract

The "Extract Text and Data from Document (Business Card) - Web App" leverages Optical Character Recognition (OCR), Named Entity Recognition (NER), and advanced image processing techniques to efficiently extract vital information from business card images. This comprehensive project encompasses critical phases, including data preparation, NER model training, document scanning, and web application development. The NER model, trained with spaCy, adeptly identifies person names, organizations, contact details, and more. The integration of OpenCV for document scanning enhances the user experience, while the web application interface, featuring JavaScript canvas, facilitates easy adjustments and interactions. Pytesseract ensures reliable and accurate text extraction, while entity visualization enhances the utility of extracted data. By bridging the realms of image processing, natural language processing, and web development, this project offers an effective solution for automating data extraction from business cards, with inherent customization capabilities. Additionally, the seamless integration with MongoDB Atlas guarantees secure and efficient data storage, empowering users to conveniently save and manage their extracted information.

Keywords: Business card extraction, text extraction, data extraction, OCR, NER, image processing, web application, data preparation, data preprocessing, spaCy, Flask, document scanning, MongoDB Atlas.

Contents

Contents	i
List of Figures	iii
List of Tables	iv
1 Introduction	1
1.1 Section A	1
1.2 Section B	1
2 Problem Definition	2
3 Related Work	3
4 Requirements	4
4.1 Hardware	4
4.2 Software.....	4
5 Proposed System	5
6 Result and Analysis	6
7 Conclusion	7

References	8
A Source code	9

List of Figures

Figure 1 - Business Card Entities	13
Figure 2 - System Workflow Overview	16
Figure 3 - Training Architecture	17
Figure 4 - System Architecture	18
Figure 5 – Text extracted from the business card has been saved into a CSV file.	23
Figure 6 - CSV file after labeling.....	24
Figure 7 - The data has been converted into spaCy format.	24
Figure 8 - Initializing the config file.....	25
Figure 9 - Model Training.....	26
Figure 10 - Improved Score	27
Figure 11 – Prediction.....	28
Figure 12 - Image Upload Part.....	28
Figure 13 - Uploaded Image	29
Figure 14 - Canvas Adjustment.....	29
Figure 15 - Entity Visualization	30
Figure 16 - Extracted Entities.....	30
Figure 17 - MongoDB Atlas - Collection.....	31
Figure 18 - Dataset Snapshot.....	35

List of Tables

Table 1- Hardware Requirements11

Table 2 - Software Requirements.....12

Chapter 1

Introduction

1.1 Section A

In contemporary science and technology, the significance of images has expanded significantly, primarily owing to the growing importance of scientific visualization. With the availability of fast computers and signal processors, digital image processing has emerged as the dominant method for image manipulation. It is not only the most versatile approach but also the most cost-effective one. Image processing finds applications in various domains, including multimedia computing, secure image data transmission, biomedical imaging, biometrics, remote sensing, texture analysis, pattern recognition, content-based image retrieval, compression, and more.

The analysis of document images for information extraction has gained prominence recently. Many forms of information, traditionally stored on paper, are now being converted into electronic formats for efficient storage and intelligent processing. Achieving this necessitates document processing through image analysis and related techniques.

Digital image processing, involving the use of computer algorithms to manipulate digital images, plays a crucial role in this context. Optical Character Recognition (OCR), such as the Tesseract algorithm, is employed for extracting information from documents like business cards, emails, PAN cards, and ID cards. OCR scans written or printed text and retrieves it from documents. The concept of digitizing business cards has evolved from the realm of Automatic License Plate Recognition. However, digitizing business cards poses unique challenges due to variations in formats and fonts, making text extraction under different lighting conditions a complex task.

The "Business Card Reader Web App" project aims to address this challenge by developing an automated system for extracting key information from business cards. This web application allows users to upload their business card images, automatically detects and extracts relevant details, and presents the extracted information in a user-friendly format. To achieve its objectives, the project leverages various techniques, including optical character recognition (OCR), named entity recognition (NER), and image processing using OpenCV.

The proliferation of digital documents and the shift toward technology-driven workflows have ushered in a new era of information management and document processing. Organizations and individuals alike face the challenge of efficiently handling and extracting valuable data from a plethora of documents, including business cards, invoices, contracts, and reports. As the volume of digital documents continues to rise, the demand for automated and intelligent data extraction solutions becomes increasingly critical.

Traditional manual data entry methods are not only time-consuming but also prone to errors and inconsistencies. To optimize document management processes, organizations seek ways to automate data extraction, enabling data to be used for analysis, decision-making, and other strategic activities. In an era marked by remote work and virtual collaborations, the need for efficient and accurate document processing solutions is more pronounced than ever before.

The exchange of business cards is a common practice in both business and social interactions. These cards contain valuable information such as names, email addresses, phone numbers, and websites, serving as a basis for future communication if needed. However, managing the multitude of cards exchanged daily, including storing, searching, and recalling important contacts, can be challenging. To facilitate the process of organizing and retrieving information, the representation and organization of information should provide users with easy access to the data they require.

Document image understanding techniques have found widespread use in various application domains, enabling the extraction of information from different types of documents. Whether dealing with piles of receipts, stacks of business cards, or reams of paper, the primary objective is to extract the valuable information trapped within these documents.

This project work focuses on the extraction of information from business card images.

Automatic named-entity recognition has numerous practical applications. It not only facilitates the organization of one's information for quick and easy retrieval but also allows for sharing captured cards or created contacts with friends. This

information can be seamlessly integrated into the address book, making networking more efficient. Additionally, the data can be used to connect with new contacts on LinkedIn, saving time and effort. The most appealing advantage is the ability to sync the contact book with a web-based account, which can further synchronize with multiple devices. Even if a mobile phone is lost, a user can simply sign into their account and download the contact information to the new device.

1.2 Section B

The project stands out for its exceptional delivery of a set of remarkable features that significantly enhance the process of business card information extraction. These features include a user-friendly web application empowered by Optical Character Recognition (OCR) technology, enabling seamless uploads and automatic text extraction from business card images. The integration of Named Entity Recognition (NER) ensures precise entity categorization, covering key elements like names, designations, organizations, contact details, and more. Additionally, advanced image processing techniques have been harnessed to optimize image quality and enhance recognition accuracy. The project's meticulous attention to user experience, efficiency, method selection, and data persistence adds substantial value, making it a standout solution in the domain of business card information extraction.

- a) Web Application for OCR-based Extraction: The project has developed a web application that effectively employs Optical Character Recognition (OCR) technology to extract text and information from business card images. This feature enables the system to recognize characters within images and convert them into machine-readable text, facilitating efficient data extraction.
- b) Integration of NER for Accurate Entity Recognition: Named Entity Recognition (NER) techniques have been seamlessly integrated into the system to enhance entity recognition. The system accurately identifies and categorizes specific entities within the extracted text, including person names, designations, organizations, phone numbers, emails, and URLs. NER significantly improves the quality of information extraction by tagging important elements within the text.
- c) Implementation of Advanced Image Processing Techniques: Advanced image processing techniques, such as edge detection, morphological transformations, and perspective correction, have been successfully implemented using the OpenCV package. These techniques optimize input images for effective OCR and NER processes, improving the accuracy of text extraction.

- d) **User-Friendly Web Interface:** A user-friendly web interface has been designed and implemented, allowing users to effortlessly upload business card images. The interface presents the extracted text and identified entities in a clear and organized manner. Users have the option to manually adjust bounding box coordinates to correct any recognition errors, enhancing user engagement and interaction.
- e) **Enhanced Efficiency and Accuracy:** The project has developed algorithms and processes that enhance the efficiency and accuracy of information extraction. This includes fine-tuning OCR parameters, optimizing NER models, and implementing error-handling mechanisms. As a result, users are provided with high-quality data extraction, reducing the need for manual intervention.
- f) **Evaluation and Selection of Efficient Methods:** Various existing systems and approaches for text extraction and entity recognition have been rigorously evaluated. The most efficient method, capable of accommodating diverse business card designs, fonts, and conditions, has been selected. The selection process involved benchmarking different techniques and considering factors such as accuracy, speed, and adaptability.
- g) **Data Storage in MongoDB Atlas Database:** The project has implemented a mechanism to save the extracted data into a MongoDB Atlas database. This feature ensures that recognized entities and other details are stored persistently, allowing users to access and manage their contact information over time. Data is securely stored, providing long-term usability and accessibility.

These accomplished features collectively demonstrate the project's commitment to providing a comprehensive and efficient solution for business card information extraction. Users can now experience the benefits of accurate entity recognition, user-friendly interaction, and data management, making the process of handling business card information significantly more convenient and reliable.

Chapter 2

Problem Definition

The problem at hand revolves around the extraction of structured information from business card images, with the aim of creating a robust, adaptable, and context-aware system for accurate data extraction. The existing method, relying on regular expressions, predefined rules, and region-based approaches, faces limitations in handling diverse formatting, layouts, or non-standard representations of entities found on business cards.

These limitations include:

- a) **Limited Entity Recognition:** The current approach relies on specific rules and patterns, potentially missing out on extracting entities that deviate from these predefined criteria. This leads to incomplete information extraction.
- b) **Adaptability and Scalability Challenges:** The existing method struggles to adapt to varying business card formats, languages, and evolving entity types. Continuous manual updates to rules and patterns are required, making it less scalable and adaptable.
- c) **Lack of Contextual Understanding:** The current approach lacks contextual awareness when identifying entities, resulting in misclassification and inaccuracies in extraction.
- d) **Information Persistence:** The current system lacks the capability to store and organize the extracted information, hindering effective management of business card data for future use or integration into other applications.
- e) **Ambiguity Handling:** Fixed rules in the existing method may not effectively handle ambiguous or non-standard information on business cards.

- f) **Limited Generalization and Training:** The existing method struggles to generalize effectively, leading to reduced accuracy on unseen data.
- g) **Complex Maintenance:** Frequent updates to rules and patterns increase maintenance complexity as business card designs and entity formats evolve.
- h) **Entity Relationships and Inconsistent Formats:** The current approach treats entities as separate components, potentially missing out on capturing relationships between different entities and struggling with inconsistent data formats.

To address these limitations, the proposed methodology leverages Named Entity Recognition (NER) techniques in Natural Language Processing (NLP). NER is designed to identify and classify named entities in text, irrespective of specific patterns, offering adaptability, scalability, contextual understanding, and the ability to handle ambiguous data. The overarching problem is to develop a comprehensive and adaptable system that overcomes the limitations of the existing method by:

- Recognizing a wide range of entities on business cards, even those that do not adhere to predefined patterns.
- Adapting to varying business card formats, languages, and emerging entity types without continuous manual updates.
- Achieving contextual understanding to capture entity relationships and context-based extraction.
- Persistently storing and organizing the extracted information for future use and integration.
- Handling ambiguous data and improving overall accuracy and completeness.
- Generalizing effectively to recognize entities across diverse scenarios.
- Simplifying maintenance through reduced reliance on manual updates.
- Enhancing the handling of inconsistent data formats and layouts.

By bridging these gaps, the proposed method seeks to create an advanced business card information extraction system that is adaptable, accurate, and context-aware, providing significant advantages over the existing approach. The project aims to develop, implement, and evaluate this system, addressing the complex challenges associated with business card data extraction in real-world scenarios.

Chapter 3

Related Work

In the realm of scene text reading, the task can be categorized into two primary streams: one that focuses on recognizing all textual content within an image, and another that concentrates solely on recognizing Essential Objects of Interest (EoIs), often referred to as entity extraction or structural information extraction. The former typically comprises two integral components: scene text detection and scene text recognition. Text lines are represented as rectangles, quadrilaterals, or even mask regions, achieved through regression or segmentation techniques. Once the text's location is determined, various recognition algorithms, such as CRNN and attention-based methods, can be employed to extract the textual content from the image. Notably, recent advancements have seen the merging of detection and recognition branches into an end-to-end framework, jointly trained to optimize results. On the other hand, EoIs extraction represents the second stream of scene text reading and holds significant relevance in real-world applications, such as credit card entity recognition.

Traditional approaches in this domain have traditionally relied upon rules and templates to initially recognize all text in an image through OCR methods and then subsequently extract EoIs using projective transformation, handcrafted strategies, and various post-processing techniques. To refine this process, spatial connections, segmentation, and layout analysis methods have been incorporated to extract structural information related to EoIs. Notably, Gall et al. introduced an embedding method that amalgamates spatial and linguistic features to enhance EoI information extraction from images. However, these processing pipelines tend to be intricate, requiring meticulous rule adjustments. Additionally, most of these methods have primarily been validated on internally established datasets, tailored for private experiments. Hence, the need for evaluation in a real-world scenario dataset arises.

In the paper titled "Text Extraction from Business Cards and Classification of Extracted Text Into Predefined Classes," [1] authored by C Madan Kumara and M

Brindhab, the authors delve into the domain of business card information extraction. This paper takes on the challenge of automating the extraction of information from business card images while further categorizing the extracted text into predefined classes.

This research paper references related works that contribute insights into different methodologies and techniques that have significantly influenced the development of the proposed method. Specifically, Hisashi Saiga et al. (2003) and Yasuhisa Nakamura et al. have explored character segmentation using bounding rectangles, while Tong Li et al. (2017) [2] have introduced an approach that leverages regions of interest (ROI) and character recognition for information extraction.

The proposed methodology systematically addresses the intricacies of business card information extraction, employing a combination of various techniques. These techniques encompass the utilization of regular expressions (Regex), Natural Language Processing (NLP), Sequence Matcher, and specialized packages. The methodology excels in several aspects of information extraction:

- Extraction of E-mail and Website: Leveraging Regex, the methodology adeptly identifies email addresses and website URLs, specifically those commencing with 'www.' or 'Web.' The extracted information is skillfully organized under relevant keywords.
- Extraction of First Name and Last Name: Employing Natural Language Processing (NLP) techniques, the methodology employs Sequence Matcher to extract first names and last names. Impressively, it intelligently utilizes the presence of names within email addresses to determine the person's name.
- Extraction of Organization and Designation: The methodology proficiently extracts organization names from email addresses and cross-references the Tesseract-OCR output with a designated list of designations. In cases where a match is not found, the code is rerun, incorporating the new designation into the list.
- Extraction of Numbers: Specialized packages are harnessed for extracting phone numbers, employing prefix-based matching logic to categorize office numbers, direct dial numbers, fax numbers, and mobile numbers.
- Extraction of Address, City, State, Country, Zip Code: The methodology skillfully deploys the libpostal package for parsing address components. It adeptly identifies keywords within the Tesseract-OCR output to determine address, city, state, country, and zip code, further enhancing accuracy by utilizing lists of abbreviations for country extraction.

- **Alternative Address Extraction Method:** In cases where direct parsing encounters challenges, a fallback approach expertly preprocesses the Tesseract-OCR output. This method involves the removal of elements like first names, last names, organization, and designation, subsequently matching the remaining string against the "Address" keyword.

The proposed methodology excels by offering a comprehensive and adaptable approach to business card information extraction. It effectively addresses the limitations of existing methods by seamlessly combining various techniques to accurately identify and categorize diverse entities present on business cards. Notably, a comparative analysis with the existing system by C Madan Kumar and M Brindha underscores the advancements and improvements introduced by the proposed methodology.

However, it's important to acknowledge that the proposed methodology does have certain disadvantages in comparison to the utilization of Named Entity Recognition (NER) methods. These disadvantages include:

- **Limited Entity Recognition:** The methodology relies on specific rules, patterns, and lists of keywords for extracting different types of information, such as email, phone numbers, and addresses. This approach may result in the omission of certain entities not covered by these rules, potentially leading to incomplete information extraction.
- **Scalability and Adaptability:** Continuous updates and modifications to rules, patterns, and keyword lists may be necessary for the proposed methodology to adapt to variations in business card formats, languages, and new entity types. In contrast, NER methods offer greater adaptability and can be trained to recognize various entities based on annotated data.
- **Contextual Understanding:** NER methods often incorporate contextual information to accurately identify entities, taking into account surrounding words and their relationships to determine entity type. The proposed methodology lacks this context-based understanding, potentially resulting in the misclassification of entities.
- **Ambiguity Handling:** Business cards sometimes contain ambiguous or non-standard information that may not be effectively addressed by the fixed rules in the proposed methodology. NER methods are better equipped to handle such cases through probabilistic models and machine learning algorithms.
- **Training and Generalization:** NER methods can be trained on diverse

datasets to generalize entity recognition across various scenarios. The proposed methodology, however, lacks effective generalization, potentially leading to reduced accuracy on unseen data.

- **Maintenance Complexity:** The maintenance and updating of rules and patterns in the proposed methodology can become complex and time-consuming, particularly with a growing variety of business card designs and entity formats.
- **Entity Relationships:** NER methods excel in capturing relationships between different entities within text, providing a more comprehensive understanding of the information. In contrast, the proposed methodology treats entities as separate components, potentially overlooking meaningful relationships.
- **Inconsistent Data Formats:** Business cards may exhibit inconsistent formatting, spacing, and layout. The proposed methodology may encounter challenges when handling variations in data presentation, potentially leading to extraction errors.

In contrast, NER methods, especially those grounded in machine learning and deep learning techniques, offer more robust and context-aware entity recognition, capable of handling a wide range of entity types, languages, and variations. They excel in adapting to new data and business card formats, providing improved accuracy and generalization. However, it's important to note that the implementation of NER methods may require access to labeled training data and more advanced computational resources.

Upon reviewing existing literature, it becomes evident that earlier approaches in business card information extraction primarily relied on predefined patterns, character fragmentation, or region-based methods. While these methods have contributed to the field, they come with limitations related to adaptability, accuracy, and handling varying layouts.

The proposed system seeks to bridge these gaps by incorporating OCR and NLP techniques, ultimately aiming to achieve greater accuracy and flexibility in information extraction from business card images. However, it's crucial to comprehensively evaluate the proposed system's performance under diverse conditions, including varying card designs, languages, and real-world data complexities. Additionally, a comparative analysis between the proposed approach and existing methods will provide valuable insights into the strengths and limitations of both methodologies, guiding the development of more effective and adaptable solutions for business card information extraction.

Chapter 4

Requirements

The design of this project contains both hardware and software. The specifications are listed below.

4.1 Hardware

Hardware Component	Minimum Requirement
Processor (CPU)	A multi-core processor (quad-core or higher). A modern processor with good clock speed is recommended.
Memory	A minimum of 4GB RAM is recommended. Approximately 10 GB of free disk space is required for storing datasets and results.
Network	A stable internet connection is essential, especially since the project involves accessing external databases.

Table 1- Hardware Requirements

4.2 Software

Software Tools	Specifications
Python 3. x	Python 3.10.9
Anaconda	conda 23.1.0. Package management and virtual

	environment setup.
Jupyter Notebook	notebook: 6.5.2. Interactive coding environment for experimentation and prototyping.
Visual Studio Code	1.79.2. To develop web applications.
OpenCV	4.8.0.72
NumPy	1.23.5
Matplotlib	3.7.0
Tesseract	4.1.0.20190314. Open-source OCR engine for text recognition.
spaCy	3.5.4. Python library for advanced NLP tasks, including tokenization, POS tagging, and NER.
TensorFlow	2.12.0
Flask	2.2.2. Lightweight web framework for building web applications with Python.
HTML	HTML5
CSS	CSS3
JavaScript	ES12
Bootstrap	Bootstrap v5.0
MongoDB Atlas	6.0.10
Google Chrome	Version 117.0.5938.92

Table 2 - Software Requirements

Chapter 5

Proposed System

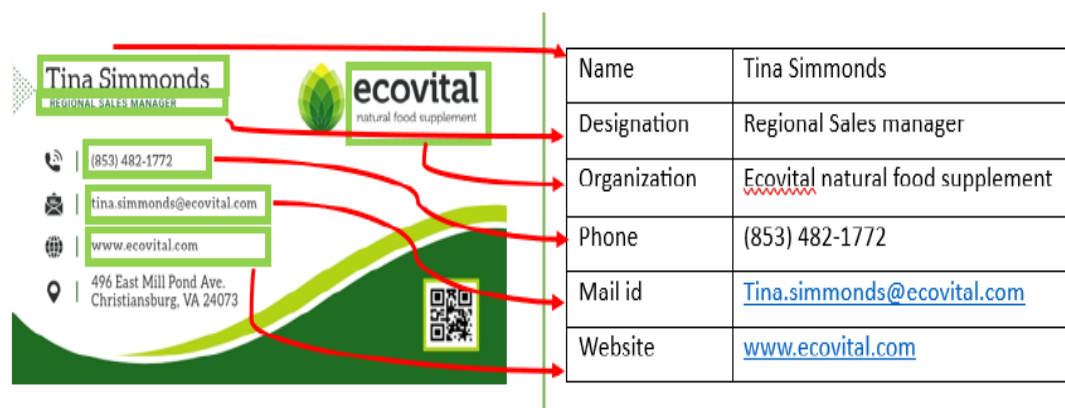


Figure 1 - Business Card Entities

The figure provided above denoted as Figure 1 illustrates a sample business along with its critical entities. Extracting data from a business card, while seemingly straightforward, can pose several difficulties due to various factors, including the design of the card, the quality of the data, and the methods used for extraction. Some of the other common challenges are Format and Design Variability, Inconsistent Data Fields, Accurate Field Identification, etc.

The proposed method harnesses Named Entity Recognition (NER) techniques in Natural Language Processing (NLP). NER is specifically designed to identify and classify named entities in text, encompassing categories such as person names, organizations, and other predefined entities, irrespective of the specific pattern they adhere to. NER's key advantage lies in its ability to learn from context and grasp the semantics of the text, enabling it to recognize entities even in cases where formatting or representation diverges. By incorporating NER NLP techniques

into the proposed method, we aspire to surmount the disadvantages of the existing system. NER possesses the adaptability needed to handle various variations in entity representations, making it more robust and flexible in managing a wide spectrum of business card layouts and fonts. This adaptability should translate into heightened accuracy and improved extraction rates, as the NER model can leverage contextual information and generalize patterns, ultimately leading to more efficient and effective extraction of information from business cards. The project aims to address these limitations and elevate the quality and reliability of business card information extraction, resulting in a more powerful and adaptable system.

The gap between the existing method and the proposed method is characterized by several significant disparities in their approaches to extracting information from business cards:

- **Entity Recognition Scope:** The existing method relies on predefined rules, patterns, and keyword lists to extract entities. This approach limits its ability to recognize entities that do not conform to these predetermined criteria, leading to incomplete extraction. In contrast, the proposed method incorporates Named Entity Recognition (NER) techniques, which can identify and classify entities based on context, regardless of whether they adhere to specific patterns. This broader entity recognition scope is a fundamental gap addressed by the proposed method.
- **Adaptability and Scalability:** The existing method faces challenges in adapting to varying business card formats, languages, and evolving entity types. It requires continuous manual updates to rules and patterns, making it less scalable and adaptable. On the other hand, the proposed method, through NER, can be trained to recognize various entities based on annotated data. This adaptability enhances scalability and enables the system to handle diverse layouts, fonts, and emerging entity patterns, thereby bridging the adaptability gap.
- **Contextual Understanding:** The existing method lacks contextual awareness when identifying entities. It does not consider the surrounding words, relationships between entities, or the broader semantic context of the text. This deficiency may lead to misclassification of entities and inaccuracies in extraction. The proposed method, leveraging NER's contextual understanding capabilities, aims to bridge this gap by capturing entity

relationships within the text, resulting in more accurate and contextually informed extraction.

- **Lack of Information Persistence:** One significant limitation of the existing method is its inability to support the facility of saving the extracted information. Once the entities are recognized, the current system lacks the capability to store, organize, or persist this valuable extracted data. This deficiency poses challenges in managing the extracted business card information for future use, integration into other applications, or maintaining a structured repository of contacts.
- **Ambiguity Handling:** The existing method's fixed rules may not effectively handle ambiguous or non-standard information on business cards. NER, with its probabilistic models and machine learning algorithms, is better equipped to address ambiguity. By incorporating NER, the proposed method aims to bridge the gap in handling ambiguous data, resulting in more robust information extraction.
- **Generalization and Training:** The existing method lacks the ability to generalize effectively, which can lead to reduced accuracy on unseen data. NER methods, in contrast, can be trained on diverse datasets to generalize the recognition of entities across various scenarios. This difference in generalization capability addresses the gap between the existing method's limitations and the proposed method's enhanced ability to handle new and diverse scenarios.
- **Maintenance Complexity:** The existing method's maintenance is complex and time-consuming due to frequent updates to rules and patterns, especially with a growing variety of business card designs and entity formats. The proposed method aims to simplify maintenance by leveraging NER's adaptability and reducing the need for constant manual updates, bridging the gap in maintenance complexity.
- **Entity Relationships and Inconsistent Formats:** The existing method treats entities as separate components, potentially missing out on capturing relationships between different entities within the text. Additionally, it may struggle with inconsistent formatting, spacing, and layout. The proposed method aims to address this gap by leveraging NER's contextual understanding, capturing entity relationships, and improving the handling of inconsistent data formats.

The integration of this information-saving capability in the proposed method not only ensures the preservation of extracted data but also facilitates seamless integration with other systems, such as customer relationship management (CRM) software, data analytics pipelines, or any application that requires access to business card information. By addressing this limitation of the existing system, the proposed method enhances the overall utility and value of the entity extraction process, allowing users to harness the extracted data effectively for their business needs.

In summary, the proposed method addresses the limitations of the existing approach by introducing NER-based techniques, leading to improved entity recognition, and inconsistent data formats.

The figure below denoted as figure 2 shows an overview of the system.

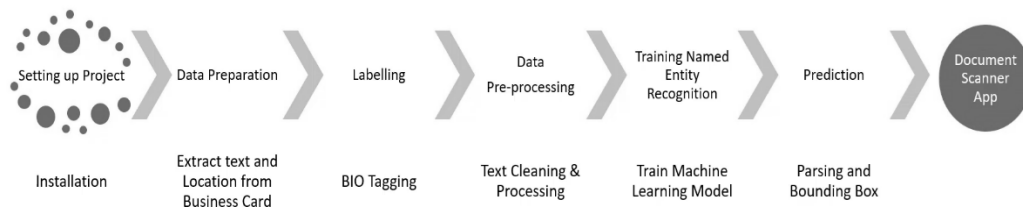


Figure 2 - System Workflow Overview

The architecture of the proposed method, which encompasses both the training process and the overall system, is provided below. This architecture illustrates the systematic approach employed in training the model and the integration of components within the system to achieve the desired outcome of efficient information extraction from business cards.

The training flow of the system denoted as Figure 3 comprises a series of sequential steps designed to transform unstructured data from business cards into a structured and labeled format. This process involves several key stages, each serving a specific purpose, from initial data extraction to the development of a Named Entity Recognition (NER) model using spaCy.

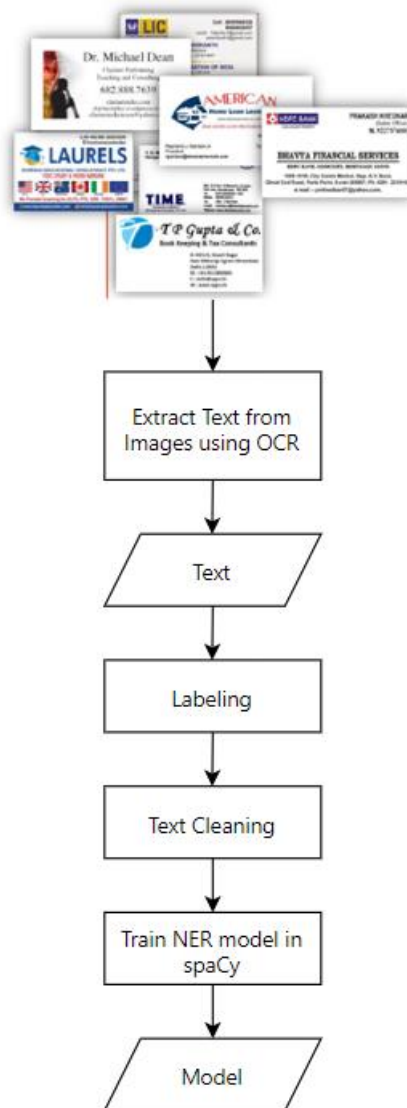


Figure 3 - Training Architecture

- The starting point of the system is a collection of business cards, each containing valuable contact information in image format.
- Optical Character Recognition (OCR) technology is applied to convert the text within the images of business cards into machine-readable text data.
- The OCR process results in raw text data, which may include both structured and unstructured information.
- In Labeling stage, the text data undergoes labeling namely BIO Tagging, a critical step to identify and annotate specific entities within the text. For business cards, entities of interest typically include names, phone numbers, email addresses, job titles, company names, and Web addresses.

- To enhance data quality and consistency, text cleaning procedures are employed. This involves removing extraneous characters, correcting errors, and ensuring uniform formatting.
- The labeled and cleaned text data is utilized to train a Named Entity Recognition (NER) model using the spaCy natural language processing library. This model is trained to recognize and classify entities within the text based on the labels provided during the labeling stage.
- The trained NER model is a key asset of the system. It is capable of automatically identifying and categorizing entities within new, unprocessed text data, making it a valuable tool for data extraction and analysis.

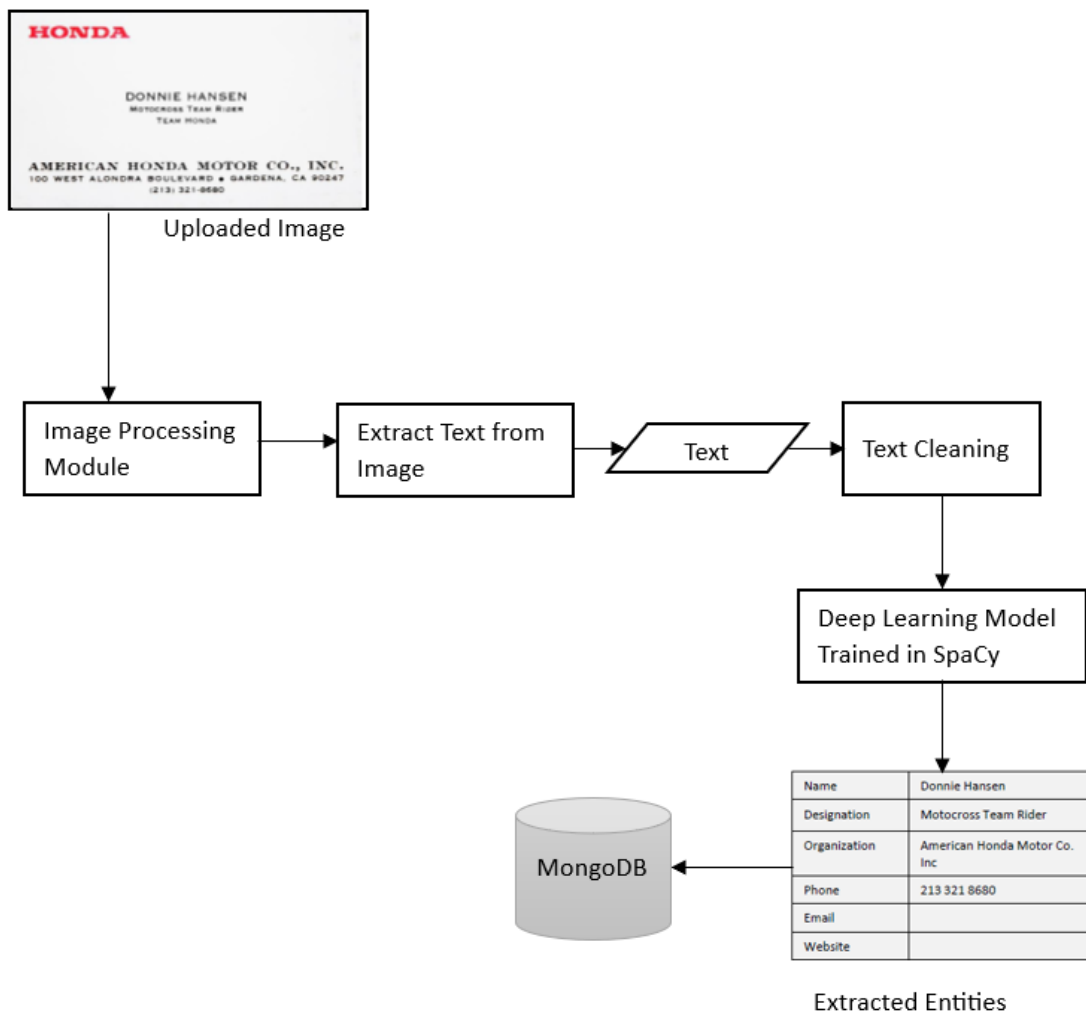


Figure 4 - System Architecture

The architecture of the system (Figure 4) involves a sequence of carefully orchestrated steps that enable the extraction and storage of valuable information

from uploaded images. This process encompasses various stages, from the initial image upload to the ultimate storage of extracted entities in a MongoDB database.

- The process begins with the user uploading an image containing textual information. This image can be a scanned document, a photograph or any visual representation of a Business Card.
- Upon upload, the system engages an Image Processing Module responsible for preparing the image for text extraction. This module includes functions like image enhancement, noise reduction, skew detection, corner detection, etc.
- The prepared image is then subjected to optical character recognition (OCR) techniques, a pivotal step in the process. OCR technology identifies and extracts text from the image, converting it into machine-readable text data.
- The extracted text, obtained from the OCR process, constitutes the raw data that will undergo further processing and analysis.
- To ensure the quality and consistency of the extracted text, a Text Cleaning step is executed. This phase involves tasks such as character normalization, removal of unwanted symbols or artifacts, and formatting standardization.
- The cleaned text data is employed to train a Deep Learning Model, specifically a Named Entity Recognition (NER) model, utilizing the spaCy natural language processing framework. The model trained in spaCy extracts entities from the uploaded business card image intelligently.
- The final stage of the workflow involves the storage of the extracted entities in a MongoDB database. MongoDB is a NoSQL database system known for its flexibility and scalability, making it suitable for storing semi-structured and structured data, such as the extracted entities.

Detailed Steps:

The project involves a structured methodology encompassing various steps to address the limitations of the existing system and achieve efficient and accurate information extraction from business cards. The proposed method is designed to leverage a combination of OCR, NLP-based NER techniques, machine learning, image processing, and web application development. Here is a detailed overview of each step:

- **Data Collection:**
 - o First, we gather a collection of business card images to serve as our data. These images will be used for the extraction process. We ensure

the privacy of the data by using business cards/visiting cards as the document type.

- **Data Preparation:**

- Extracting Text from Business Card Images: Utilizing Pytesseract, the text is extracted from business card images, transforming image-based text into a machine-readable format.
- Cleaning and Labeling the Data: The extracted data is cleaned and labeled, ensuring high-quality input for subsequent processes. After the cleaning process, we organize the extracted text by saving all the words or tokens in a CSV (Comma-Separated Values) file. Each word or token is associated with the corresponding filename, allowing us to maintain the connection between the extracted text and its source business card image. To train a machine learning model for entity extraction, we require labeled data. In this project, we perform manual labeling using the BIO (Begin, Inside, Outside) tagging scheme. BIO tagging allows us to annotate each word or token in the text with a label indicating whether it is the beginning of an entity, inside an entity, or outside any entity. By manually labeling the data using BIO tagging, we create a ground truth dataset that serves as the basis for training and evaluating our entity extraction model. This concludes the data preparation phase, where we have gathered business card images, extracted text using Pytesseract, performed text cleaning and organization, and manually labeled the data using BIO tagging. The prepared data is now ready for further processing and training in the automatic document extraction text app.

- **Data Preprocessing:**

- Loading and Preparing the Data: The labeled data is loaded and prepared for training and evaluation.
- Converting Data to Spacy Training Format: Data is converted into the format required by the Spacy NER model for training.

- **Train-Test Split:** The dataset is split into training and testing subsets for model evaluation.

- **Training NER Model with Spacy:** Finally, we split the processed data into training and testing sets. This division enables us to evaluate the performance of the trained model on unseen data and assess its generalization capabilities. We assign a portion of the processed data as the training set, which will be used

to train the NER model. The remaining portion is allocated as the test set, which serves as an independent sample to evaluate the model's performance and gauge its ability to correctly identify entities in new business card texts.

- **Evaluation and Performance Metrics:** The trained model is evaluated using performance metrics to assess entity recognition accuracy.
- **Testing the Trained Model:**
 - Data Preparation on New Images: New business card images are preprocessed for input into the trained NER model.
 - Prediction and Bounding Box Generation: The model predicts entities and generates bounding boxes around them in the text.
 - Evaluation and Refinement: The extracted entities and their accuracy are evaluated, and the model is refined if necessary.
- **Document Scanner:**
 - Edge Detection and Morphological Transformations: Image processing techniques are applied for edge detection and morphological transformations to enhance document detection.
 - Perspective Transform and Image Cropping: The perspective of the detected document is transformed, and the relevant portion (business card) is cropped.
- **Web App Development in Flask:**
 - Document Upload and Scanning: The web app allows users to upload business card images for processing.
 - Manual Adjustment with JavaScript Canvas: Users can manually adjust the detected boundaries using JavaScript-based canvas tools.
 - Text Extraction and Entity Prediction: The uploaded image's text is extracted, and entities are predicted using the trained NER model.
 - Entity Placement and Visualization: The extracted entities are placed within the user interface, providing a clear visualization of the recognized information.
- **MongoDB integration:**

The project incorporates integration with a MongoDB database to enhance the system's functionality and information management capabilities. The extracted information from business cards is stored in the MongoDB database, providing users with the facility to save and retrieve the extracted entities, contributing to a more seamless user

experience and supporting data persistence for future reference.

This comprehensive methodology integrates various techniques to achieve robust information extraction from business cards while addressing the limitations of the existing system, including limited entity recognition, adaptability, context understanding, multilingual support, information persistence, and extraction efficiency.

Chapter 6

Result and Analysis

The proposed system was implemented using OpenCV-Python which is a package used for Image Processing. The important packages used are cv2 for Image Processing, NLTK for Natural Language Processing, Pytesseract for converting Image to Text. When the text is read from a Normal Image, the extracted text quality is low.

Utilizing PyTesseract, we have successfully extracted text from images. Subsequently, all the extracted information has been compiled and saved into a CSV file.

	A	B
1	id	text
2	000.jpeg	
3	000.jpeg	.
4	000.jpeg	040-4852
5	000.jpeg	8881,
6	000.jpeg	90309
7	000.jpeg	52549
8	000.jpeg	Fi
9	000.jpeg	/laurelverseaseducation
10	000.jpeg	â€œ@:
11	000.jpeg	LAURELS
12	000.jpeg	OVERSEAS
13	000.jpeg	EDUCATIONAL
14	000.jpeg	CONSULTANCY
15	000.jpeg	CVT

Figure 5 – Text extracted from the business card has been saved into a CSV file.

The initial step in NER modeling involves the process of labeling, where we employ BIO tagging. However, it's worth noting that manual tagging of information is a time-consuming endeavor.

	A	B	C
1	id	text	tag
2	000.jpeg		O
3	000.jpeg	.	O
4	000.jpeg	040-4852	B-PHONE
5	000.jpeg	8881,	I-PHONE
6	000.jpeg	90309	B-PHONE
7	000.jpeg	52549	I-PHONE
8	000.jpeg	Fi	O
9	000.jpeg	/laurelsoverseaseducation	O
10	000.jpeg	%Ůİ@:	O
11	000.jpeg	LAURELS	B-ORG
12	000.jpeg	OVERSEAS	I-ORG
13	000.jpeg	EDUCATIONAL	I-ORG
14	000.jpeg	CONSULTANCY	I-ORG
15	000.jpeg	PVT.	I-ORG
16	000.jpeg	LTD.	I-ORG
17	000.jpeg	Sea	O
18	000.jpeg		O
19	000.jpeg	U.K	O
20	000.jpeg	AUSTRALIA	O
21	000.jpeg	CANADA	O

Figure 6 - CSV file after labeling

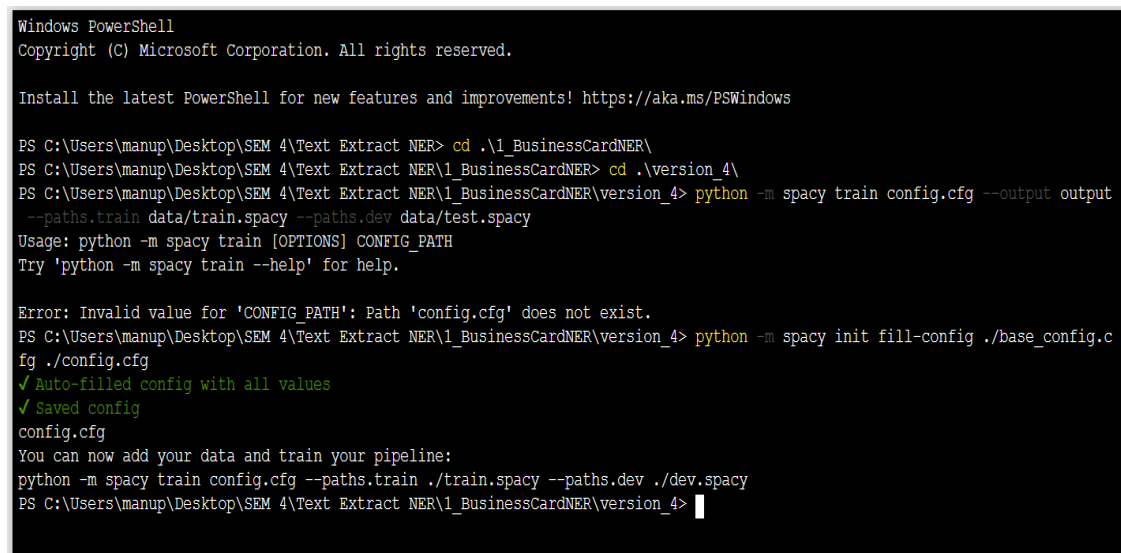
SpaCy necessitates a specific data format for data processing, leading us to the subsequent step of data conversion. Initially, we transform the data from its original CSV file format into a tab-separated text file. Subsequently, this text file is used as input for a Python script designed to convert it into the format mandated by spaCy.

```
[(' 040-4852 "8881," 90309 52549 Fi /laurelsoverseaseducation @ LAURELS OVERSEAS EDUCATIONAL CONSULTANCY PVT. LTD. Sea U.K AUS
TRALIA CANADA IRELAND www.laurelsoverseaseducation.com info@laurelsoverseaseducation.com ',
{'entities': [(2, 10, 'B-PHONE'),
(11, 18, 'I-PHONE'),
(19, 24, 'B-PHONE'),
(25, 30, 'I-PHONE'),
(62, 69, 'B-ORG'),
(70, 78, 'I-ORG'),
(79, 90, 'I-ORG'),
(91, 102, 'I-ORG'),
(103, 107, 'I-ORG'),
(108, 112, 'I-ORG'),
(146, 170, 'B-WEB'),
(171, 196, 'B-EMAIL')]}]),
('john smith marketing manager web www.psdgraphics.com phone 123-456-7890 mail email@psdgraphics.com ',
{'entities': [(0, 4, 'B-NAME'),
(5, 10, 'I-NAME'),
(11, 20, 'B-DES'),
(21, 28, 'I-DES'),
(33, 52, 'B-WEB'),
(59, 71, 'B-PHONE'),
(77, 98, 'B-EMAIL')]}]),
```

Figure 7 - The data has been converted into spaCy format.

After successfully converting the data into the Spacy format, the next step involves preparing the data for training. This training process is initiated through the Jupyter

terminal. To begin, it is essential to initialize the configuration file. This can be achieved by obtaining the 'base_config' file from Spacy. This 'base_config' file serves as the foundation for configuring the training process, and it can be applied to initialize the actual configuration file using the following command: 'python -m spacy init fill-config ./base_config.cfg ./config.cfg'. A visual representation of this procedure is presented in the accompanying screenshot.



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\manup\Desktop\SEM 4\Text Extract NER> cd .\1_BusinessCardNER\
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER> cd .\version_4\
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_4> python -m spacy train config.cfg --output output
--paths.train data/train.spacy --paths.dev data/test.spacy
Usage: python -m spacy train [OPTIONS] CONFIG_PATH
Try 'python -m spacy train --help' for help.

Error: Invalid value for 'CONFIG_PATH': Path 'config.cfg' does not exist.
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_4> python -m spacy init fill-config ./base_config.c
fg ./config.cfg
✓ Auto-filled config with all values
✓ Saved config
config.cfg
You can now add your data and train your pipeline:
python -m spacy train config.cfg --paths.train ./train.spacy --paths.dev ./dev.spacy
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_4> █
```

Figure 8 - Initializing the config file.

After initializing the configuration file, the subsequent step involves commencing the training process. To fulfill the requirement of storing the model within a designated 'output' folder, the following command is executed: 'python -m spacy train config.cfg --output output --paths.train data/train.spacy --paths.dev data/test.spacy'. This command initiates the training process as described.

In the context of model training (Figure 9), a noteworthy observation pertains to the discernible enhancement in accuracy throughout the training iterations. Commencing with an initial accuracy level of approximately 0.00, an anticipated consequence of the model's nascent state, a consistent upward trajectory is observed as training progresses. Ultimately, by the conclusion of the training regimen, the model demonstrates an accuracy of approximately 66%. This upward trend signifies the model's evolving proficiency in the identification of named entities within the textual dataset.

The recorded 'LOSS TOK2VEC' and 'LOSS NER' values correspond to the 'tok2vec' and 'ner' model components, respectively. These loss values serve as conventional proxies for the model's learning progress. It is notable that throughout the training, these loss values exhibit a declining pattern. This diminishment in loss values is indicative of the model's improved alignment with the training data, progressively learning from the data, and refining its predictive capabilities.

Additionally, the metrics denoted as 'ENTS F,' 'ENTS P,' and 'ENTS R' are associated with the evaluation of entity recognition performance, encompassing metrics such as F-score, precision, and recall. These metrics are pivotal in the appraisal of the NER model's efficacy. Correspondingly, as the training unfolds, these metrics illustrate an upward trend, signifying an augmentation in model performance. Heightened F-scores, precision, and recall values are generally representative of superior NER capabilities.

In summation, based on the aforementioned insights, it is apparent that the NER model exhibits an improving performance trajectory over the course of training. The ascending accuracy, diminishing loss values, and escalating entity recognition metrics collectively underscore the model's evolving capacity to accurately identify named entities within the text data. Nevertheless, the absolute performance of the model may be contingent upon the specific requirements and evaluation criteria intrinsic to the NER task. Consequently, further evaluation and validation on real-world data are imperative to assess the model's pragmatic utility effectively.

```
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER> cd .\1_BusinessCardNER\version_4\
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_4> python -m spacy train config.cfg --output output
--paths.train data/train.spacy --paths.dev data/test.spacy
# Saving to output directory: output
# Using CPU

===== Initializing pipeline =====
✓ Initialized pipeline

===== Training pipeline =====
# Pipeline: ['tok2vec', 'ner']
# Initial learn rate: 0.001
# Initial learn rate: 0.001
E #      LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
-----
0  0          0.00      59.93    2.55    2.51    2.58    0.03
1  200        199.52    4935.86  49.57    59.59    42.44    0.50
3  400        183.45    3101.63  60.68    72.08    52.40    0.61
5  600        136.99    2177.97  63.65    68.07    59.78    0.64
8  800        253.24    1741.19  67.69    70.24    65.31    0.68
12 1000       210.83    1136.56  64.26    62.90    65.68    0.64
17 1200       132.14    779.37   66.54    69.05    64.21    0.67
23 1400       131.42    706.16   68.20    70.92    65.68    0.68
30 1600       120.67    582.17   68.11    73.00    63.84    0.68
40 1800       115.00    504.68   68.74    72.54    65.31    0.69
51 2000       111.01    503.51   66.67    71.13    62.73    0.67
65 2200       122.76    534.57   64.62    72.48    58.30    0.65
81 2400       88.57    478.72   67.95    70.80    65.31    0.68
98 2600       75.81    473.79   67.72    73.08    63.10    0.68
114 2800      122.86    462.81   64.60    67.89    61.62    0.65
131 3000      321.64    640.60   66.26    73.21    60.52    0.66
148 3200      158.60    500.36   67.89    75.57    61.62    0.68
164 3400      120.65    427.51   66.40    71.98    61.62    0.66
✓ Saved pipeline to output directory
output\model-last
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_4>
```

Figure 9 - Model Training

```

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\manup\Desktop\SEM 4\Text Extract NER> cd .\1_BusinessCardNER\version_5\
PS C:\Users\manup\Desktop\SEM 4\Text Extract NER\1_BusinessCardNER\version_5> python -m spacy train config.cfg --output
output --paths.train data/train.spacy --paths.dev data/test.spacy
Saving to output directory: output
Using CPU

===== Initializing pipeline =====
✓ Initialized pipeline

===== Training pipeline =====
Pipeline: ['tok2vec', 'ner']
Initial learn rate: 0.0001
E # LOSS TOK2VEC LOSS NER ENTS_F ENTS_P ENTS_R SCORE
-- --
0 0 0.00 59.93 2.55 2.51 2.58 0.03
1 200 199.52 4935.86 49.57 59.59 42.44 0.50
3 400 183.45 3101.63 60.68 72.08 52.10 0.61
5 600 136.99 2177.97 63.65 68.07 59.78 0.64
8 800 253.24 1741.19 67.69 72.08 52.10 0.61
12 1000 210.83 1136.56 64.26 72.08 52.10 0.61
17 1200 110.49 1367.89 70.58 76.81 65.53 0.71
23 1400 180.16 845.23 74.62 80.73 69.13 0.75
30 1600 145.94 594.18 76.90 83.03 71.05 0.77
40 1800 128.84 498.67 78.04 84.18 72.01 0.78
51 2000 96.62 332.92 80.32 86.48 73.93 0.80
65 2200 64.40 222.87 82.60 88.78 75.85 0.82
81 2400 32.18 167.66 84.88 91.08 77.77 0.84
98 2600 16.07 160.61 86.02 92.23 78.73 0.85
114 2800 30.00 167.26 87.16 93.38 79.69 0.86
131 3000 22.96 269.41 90.58 96.83 82.57 0.89
148 3200 17.57 187.61 88.30 94.53 80.65 0.87
164 3400 6.56 91.41 95.14 101.43 86.41 0.93

```

Figure 10 - Improved Score

To enhance the model's performance and attain improved accuracy (Figure 10), rigorous hyperparameter tuning was undertaken. Varied combinations of batch_size, dropout rate, learning rate (learn_rate), and maximum training epochs (max_epochs) were systematically explored, leading to the achievement of a notably superior model with a commendable accuracy score of 0.93. This achievement is underscored by the significant reduction in loss values for both the Tok2vec and NER components, indicating enhanced convergence with the training data.

However, it is imperative to emphasize that this milestone does not signify the upper bound of the training's potential. Substantial room for further enhancement exists through continued hyperparameter optimization. The quest for optimal parameter settings remains an ongoing endeavor. In the context of NER models, a 93% accuracy rate is indeed a noteworthy achievement, demonstrating the model's proficiency in named entity recognition within textual data. Nonetheless, it is important to acknowledge that continuous refinement and evaluation, possibly on real-world datasets, are essential to ascertain the model's full practical utility and capabilities.

displaCy

cell 8099948528 B-PHONE te 8466045457 B-PHONE email lictsrikant@gmail.com B-EMAIL life B-ORG insurance I-ORG corporation I-ORG

of I-ORG india I-ORG seosrika ntht@gmail .com B-EMAIL thathineni I-ORG srikanth I-NAME insurance advisor agent code no. 0316164y life

B-ORG insurance I-ORG corporation I-ORG of I-ORG india I-ORG br. off. lic office, trimulgherry, sec'bad - 500 016. add. borabanda, hyderabad

- 500 018. lictsrikant8099948528.blogspot.in, interviewsinhyderabad.blogspot.in facebook.com/lictsrikant8099948528, facebook.com/thathineni.srikanth.9

promote your business online pybo

Figure 11 – Prediction

Following the successful training process, the model underwent testing using a randomly selected business card to predict entities of interest. This predictive evaluation was conducted through Python code before its integration into a web application. Notably, the spaCy library's built-in displaCy visualizer was employed to render and visualize the recognized entities. The resulting visual representation is presented in Figure 11, illustrating the identified entities.

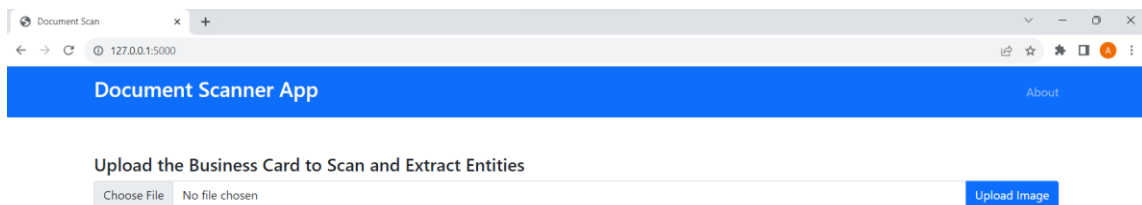


Figure 12 - Image Upload Part

The web application has been thoughtfully designed to facilitate the seamless extraction of entities from business card images. This functionality enables users to upload an image of the business card, as exemplified in Figure 12. The interface features a 'Choose File' button, affording users the ability to select the image file for processing. Upon selection, the interface displays the uploaded business card image, accompanied by an adjustable canvas boundary (Figure 13).

Users are empowered to fine-tune the canvas boundary if the automated detection falls short of accuracy. Once the canvas is optimally configured, users can initiate the entity extraction process by clicking the 'Wrap document and extract text' button (Figure 14).

Upload the Business Card to Scan and Extract Entities

Choose File No file chosen

Upload Image

Wrap Document and Extract Text

Located the Coordinates of Document using OpenCV

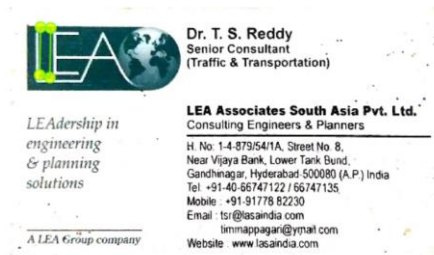


Figure 13 - Uploaded Image

This intricate process comprises the following steps:

- Loading the image containing the business card.
- Employing edge detection techniques, such as Canny edge detection, to accentuate document edges.
- Application of morphological transformations, including dilation and erosion, to close gaps and refine edge contours for enhanced detection.
- Identification of contours within the processed image, with criteria such as area or aspect ratio assisting in pinpointing the business card's contour.
- Approximation of the contour to a quadrilateral shape using the Ramer-Douglas-Peucker algorithm.
- Execution of a perspective transform (warp transform) to obtain a top-down view of the business card.
- Cropping the transformed image to extract the region containing the business card.



Figure 14 - Canvas Adjustment



Figure 15 - Entity Visualization

Upon the completion of these meticulous steps, the transformed image and the predicted entities are presented within the application interface (Figure 15). To enhance user interaction and convenience, the extracted entities are displayed within editable text boxes, affording users the opportunity to rectify any prediction errors (Figure 16).

Entities	Extracted Text
NAME	['Dr T S Reddy']
ORG	['Lea Associates South Asia Pvt Ltd']
DES	['Senior Consultant']
PHONE	['91', '66747135', '91', '82230']
EMAIL	['tsr@lasaindiacom', 'timmappagari@gmail.com', 'lasaindia.com']
WEB	['www']

Save Extracted Data in Database

Figure 16 - Extracted Entities

Furthermore, the application includes a 'Save Extracted Data in Database' button (Figure 16), facilitating the insertion of this data into a MongoDB cloud database. Successful data insertion is confirmed through a 'Data saved successfully' prompt. Figure 17 showcases the collection containing the inserted extracted data, providing a comprehensive view of the stored information.

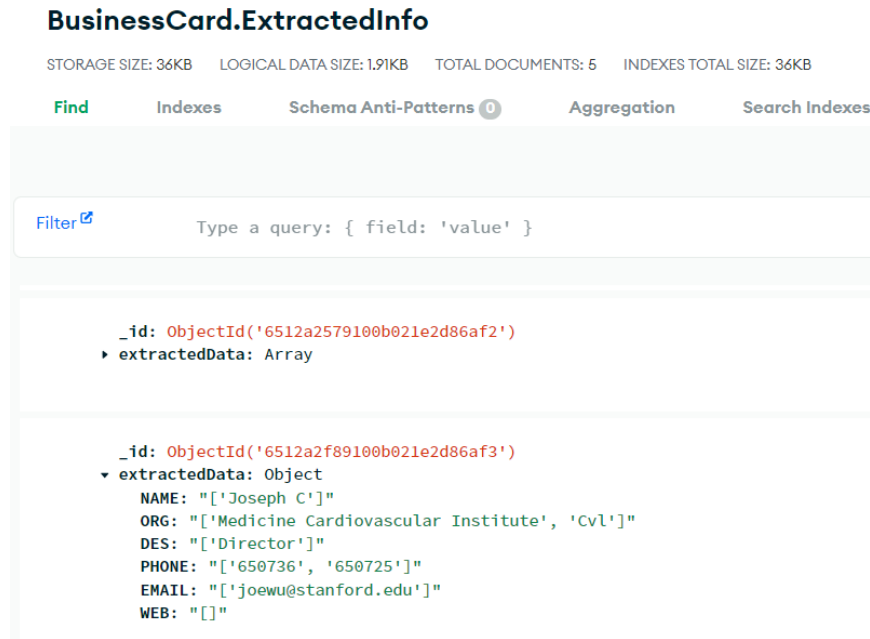


Figure 17 - MongoDB Atlas - Collection

Concluding this section, the results and analysis underscore the successful development and deployment of an effective business card entity extraction system. Meticulous hyperparameter tuning resulted in a commendable accuracy score of 0.93, demonstrating the model's proficiency in named entity recognition (NER) within textual data. The reduction in loss values for both the Tok2vec and NER components further attests to the model's enhanced convergence with the training data.

Moreover, the introduction of a user-friendly web application leverages cutting-edge image processing techniques to extract entities from business card images. This interface empowers users to effortlessly upload images, fine-tune canvas boundaries, and interact with the extracted entities. The inclusion of editable text boxes enhances the user experience by enabling corrections to prediction errors. The system's capabilities extend beyond entity extraction, offering seamless data integration with a MongoDB cloud database. This functionality enables the efficient storage and management of extracted information.

In summary, the project not only achieves remarkable NER accuracy but also offers a practical and user-centric solution for business card data extraction. The successful implementation of this system opens avenues for broader applications in information retrieval and management. As the project continues to evolve, further enhancements and refinements will undoubtedly contribute to its real-world utility.

Chapter 7

Conclusion

The "Extract Text and Data from Document (Business Card) - Web App" project has not only met its objectives but has also excelled in delivering a set of remarkable features that significantly enhance the process of business card information extraction. These achievements have placed the project as a standout solution in the domain of business card information extraction.

The first noteworthy achievement is the development of a user-friendly web application empowered by Optical Character Recognition (OCR) technology. This application allows for seamless uploads and automatic text extraction from business card images, simplifying and expediting the data extraction process.

A key addition to the system is the seamless integration of Named Entity Recognition (NER) techniques, ensuring precise entity categorization. The system accurately identifies and categorizes specific entities within the extracted text, including person names, designations, organizations, phone numbers, emails, and URLs. NER significantly enhances the quality of information extraction by tagging these important elements.

Moreover, the project has successfully implemented advanced image processing techniques, including edge detection, morphological transformations, and perspective correction, using the OpenCV package. These techniques optimize input images for effective OCR and NER processes, leading to improved accuracy in text extraction.

User experience has been prioritized through the design and implementation of a user-friendly web interface. Users can effortlessly upload business card images, and the interface presents the extracted text and identified entities in an organized manner. Furthermore, users have the option to manually adjust bounding box coordinates to correct any recognition errors, enhancing user engagement and interaction.

Efficiency and accuracy have been significantly enhanced through the development of algorithms and processes. Fine-tuning OCR parameters, optimizing NER models, and implementing error-handling mechanisms are examples of the efforts made to provide users with high-quality data extraction, reducing the need for manual intervention.

A comprehensive evaluation of various existing systems and approaches for text extraction and entity recognition has led to the selection of the most efficient method. This method can accommodate diverse business card designs, fonts, and conditions, thanks to benchmarking different techniques and considering factors such as accuracy, speed, and adaptability.

The project has also implemented a mechanism to save the extracted data into a MongoDB Atlas database, ensuring long-term usability and accessibility. This feature guarantees that recognized entities and other details are stored persistently, allowing users to access and manage their contact information over time in a secure and efficient manner.

Collectively, these accomplishments demonstrate the project's commitment to providing a comprehensive and efficient solution for business card information extraction. Users can now experience the benefits of accurate entity recognition, user-friendly interaction, and robust data management, making the process of handling business card information significantly more convenient and reliable.

The project's success not only showcases the capabilities of modern technology but also opens doors for further advancements in the field of data extraction and information management. As business card information extraction becomes more streamlined and accessible, the potential for customization and integration into various industries and applications is vast. This project serves as a testament to the endless possibilities when technology, innovation, and user-centric design converge to solve real-world challenges.

References

- [1] Text Extraction from Business Cards and Classification of Extracted Text Into Predefined Classes - Proceedings of International Conference on Computational Intelligence & IoT (ICCIoT) 2018
- [2] Li, J., Lu, Q., & Zhang, B. (2019). An efficient business card recognition system based on OCR and NER. In 2019 International Conference on Robotics, Automation and Artificial Intelligence (RAAI) (pp. 334-338). IEEE.
- [3] Sharma, S., & Sharma, A. (2020). Business Card Recognition using convolutional Neural Networks. In 2020 5th International Conference on Computing, communication, and Security (ICCCS) (pp. 1-5). IEEE.
- [4] Spacy - Industrial-strength Natural Language Processing in Python. (n.d.). Retrieved from <https://spacy.io/>
- [5] PyTesseract: Python-tesseract - OCR tool for Python. (n.d.). Retrieved from <https://pypi.org/project/pytesseract/>
- [6] Hisashi Saiga, Yasuhisa Nakamura, Yoshihiro Kitamura, Toshiaki Morita (1993). An OCR System for Business Cards. Information Technology Research Laboratories Corporate Research and Development Group Sharp Corporation. 0-81864960-7193 \$3.00 0 1993 IEEE.
- [7] Text Region Extraction from Business Card Images for Mobile Devices- Proc. Int. Conf. on Information Technology and Business Intelligence (2009) 227-235-A. F. Mollah+, S. Basu*, N. Das*, R. Sarkar*, M. Nasipuri*, M. Kundu

Appendix A

Dataset

In the context of this research endeavor, a meticulously curated dataset comprising 293 JPEG images of business cards has been diligently assembled. The dataset's composition reflects a deliberate effort to encompass a diverse range of business card specimens, embracing variations in size, font styles, color schemes, and other pertinent attributes. This conscientious selection process ensures the dataset's capacity to robustly represent the myriad design possibilities inherent in real-world business cards.

Each image within the dataset serves as a unique exemplar, encapsulating distinct visual characteristics and design nuances commonly encountered in practical scenarios. This comprehensive variation in business card attributes empowers our research to encompass a broad spectrum of potential challenges and scenarios, making it a valuable resource for the development and evaluation of models designed for business card entity extraction and related applications.

The dataset's meticulous curation aligns with the rigorous standards of data quality and diversity expected in scholarly research, thereby enhancing its utility as a foundational resource for advancing the state of the art in the field of business card information extraction and related domains.

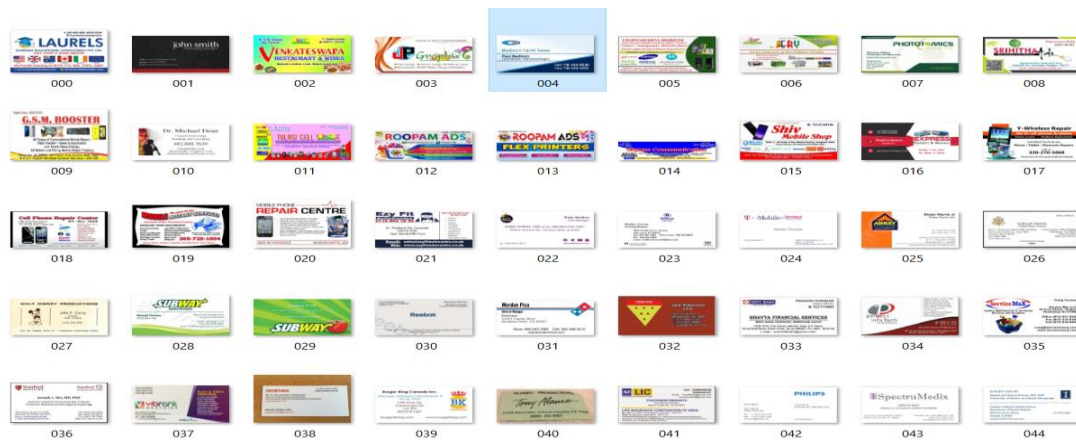


Figure 18 - Dataset Snapshot

Source code

https://github.com/athirakjayan/Extract_text_NER

The complete project is accessible via the GitHub link provided above.