University of Prince Mugrin
College of Computer and Cyber Sciences
Department of Artificial Intelligence

AI312- Natural Language Processing

Course Project – Semester II (Spring 2024)

Resume Analysis: Information Extraction

Team Members:

Rasha Ashawa | 4110496

Salwa Shamma | 4010405

Sana Shamma | 4010404

Nafisah Shams | 4010434

Instructor:

Dr. Syed Bukhari

May 12, 2024

TABLE OF CONTENTS

Abstract

This report highlights the development and evaluation of a resume analysis system, particularly in information extraction. The resume analysis system utilizes one of the Natural Language Processing (NLP) tasks, which is Named Entity Recognition (NER), to automatically extract and categorize information from resumes, such as personal information, educational background, work experience, and skills. The aim is to facilitate human resources processes by automatically analyzing different sections of resumes.

Keywords: Resume Analysis, Information Extraction, Natural Language Processing (NLP), Named Entity Recognition (NER).

CHAPTER 1: INTRODUCTION

1.1 Overview

Over the last 10 years, there has been a paradigm shift in the recruitment field, moving from traditional job fairs to modern e-recruiting website platforms. Nowadays, well-known recruiting platforms such as LinkedIn and Monster publish numerous resumes daily, also enabling job seekers to automatically create their own resumes. As a result, this shift has placed a burden on human resources departments, as they now have to read and process a significant volume of job applications. Therefore, this has motivated us to address this issue and automate these routine tasks by developing a resume analysis system that uses the magic of NLP techniques to extract and categorize resume data.

CHAPTER 2: METHODOLOGY

2.1 Selected Method

The resume analysis system utilizes one tool from the Natural Language Processing Toolkit, which is Named Entity Recognition (NER), to extract information from resumes. Instead of building our model from scratch, we decided to build our model on an already pre-trained model provided by SpaCy called 'ner', which detects entities such as Cardinal, Date, Event, FAC, GPE, Language, Law, Location, Money, NORP, Ordinal, Organization, Percent, Person, Product, Quantity, Time, Work_of_Art, and extends to detect some entities that are related to resume analysis context, such as skills, experiences, college, and so on.

2.2 Feature Vectors

Our feature vectors have two tuples; the first tuple is full data, which is the textual content of the resumes, and the second tuple is the label on which the model will be trained in the resume analysis context. The label is a dictionary in our case called entities; its values are an array whose elements are tuples. The second index of the tuple is the label for named entity recognition, the zero index is the location of the first characters of the value of named entity recognition, and the first index is the last location of the value of named entity recognition. Figure 2.1 illustrates the train dataset structure.

Figure 2.1 - Train Dataset Structure

2.3 Pseudocode

```
BEGIN
# Pre-Training Step
Load dataset
Split data into train dataset and test dataset
Initialize pre-trained model
# Train The Model
While num_iteration < num_epochs:
        Shuffle train dataset
        For text, annotations in train_data:
                Feed model with train data and train parameters
# Test The Model
For text, annotations in test_data:
        Feed model with test dataset
        Calculate the accuracy
Display the accuracy
If model accuracy is acceptable:
        Save the model
END
```

2.4 Flow-chart

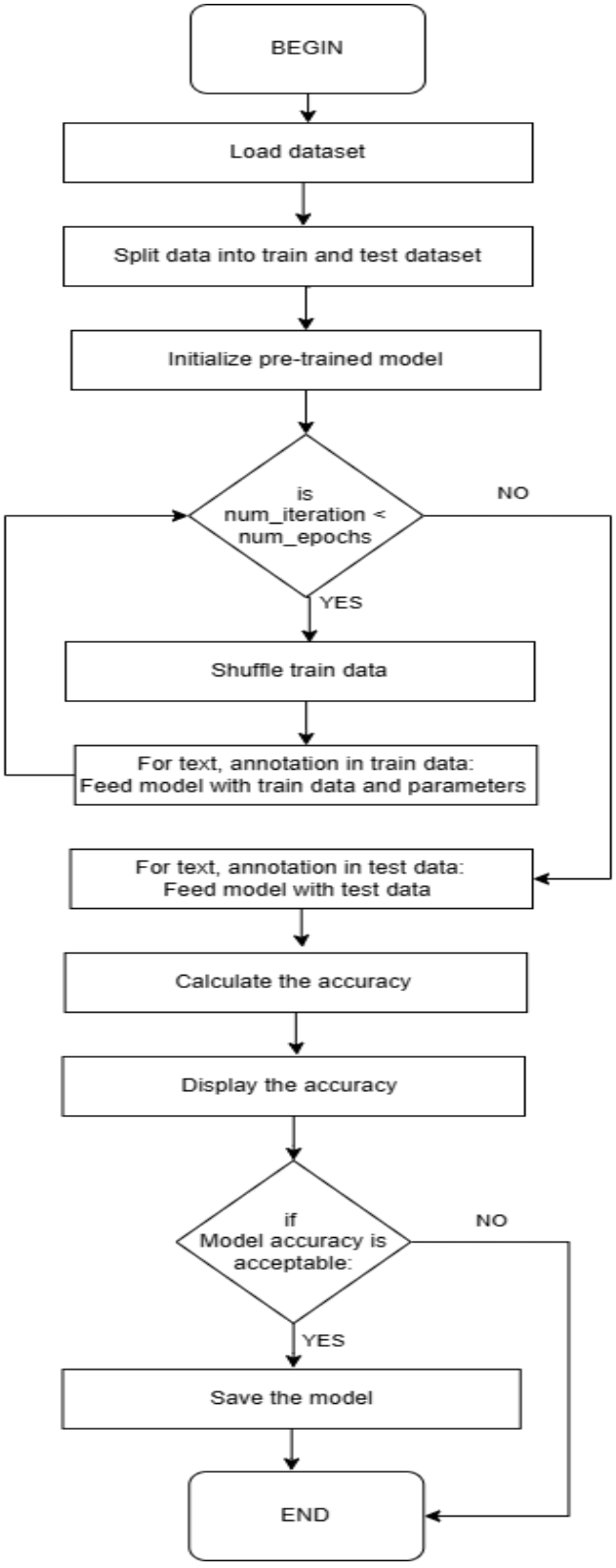

Figure 2.1 - Flow Chart of the resume analysis system

CHAPTER 3: RESULTS

3.1 Test Dataset

The system was tested on a dataset of 10 resumes, achieving an accuracy of 44% in correctly categorizing different sections of the resume. We achieved this result after training the model for 250 epochs on 190 training samples. The results are presented in graphical format to illustrate the system's performance across various categories, as shown in Figure 3.1.

```
NAME                          - kimaya sonawane
LOCATION                      - Thane
EMAIL ADDRESS                 - indeed.com/r/kimaya- sonawane/1f27a18d2e4b1948
DESIGNATION                   - Technical Support Engineer
COMPANIES WORKED AT           - SAP -
LOCATION                      - Thane
GRADUATION YEAR               - 2016
DEGREE                        - BE in computer science
COLLEGE NAME                  - SSVPS's Late B. S. Deore College of Engineering
GRADUATION YEAR               - 2016
SKILLS                        - network engineers, Networking, CCNA, knowledge of Active Directory, DHCP, DNS , Troubleshooting
and fix Network related issues (2 years)
GRADUATION YEAR               - 2016
```

Figure 3.1 - System Outputs Formats

```
Precision for the test set: 0.43617021276595747
Recall: 0.2645161290322581
F-Score: 0.3293172690763052
```

Figure 3.2 - System Accurancy

3.2 Real-Time Test

The system has the capability to receive a CV as input and provide real-time analysis to extract information. We achieved that by first converting the input file to PDF format using PyMuPDF, then we read the CV line as string text. Then what we read from the CV lines was entered into our trained model. Figure 3.3 shows the input data, while Figure 3.4 shows the system outputs.

Figure 3.3 - Input Sample

```
NAME                    - RAJESH
LOCATION                - Hyderabad
GRADUATION YEAR         - 2015
COLLEGE NAME            - Borcelle University Bachelor of Business Management
```

Figure 3.4 - System Output

CHAPTER 4: CONCLUSION & OUTLOOK

4.1 Conclusion

In summary, we applied our knowledge of NLP to our project. We began by describing our methodology that we followed, then implemented one of the NLP tasks which is Named Entity Recognition (NER), to be able to develop this application. Throughout the project, we utilized the techniques and concepts we learned in our lectures, improving our understanding through hands-on application. Overall, this project allowed us to gain practical experience and reinforce our understanding of the course materials.

4.2 Future Work

It is crucial to declare the limitations of our current study and what the next steps are to address them in future work. Such work will focus on improving the model's accuracy by training it on a larger dataset with a wider range of information categories to make it a more effective system. In addition, we will integrate it with RNNs to gain more context and better extract information. The current system focuses on the English language; in the future, we aim to support the Arabic language as well. Furthermore, we will add another feature to provide advice and suggest keywords that would make a CV more attractive and improve it.

References

1. Lakhani, G. (n.d.). Resume and CV Summarization. KGP Talkie. Retrieved from https://kgptalkie.com/resume-and-cv-summarization/
2. KGP Talkie. (n.d.). [NLP Tutorial 16 - CV and Resume Parsing with Custom NER Training with SpaCy]. YouTube. https://www.youtube.com/watch?v=WpaioLNsoGI&t=1073s
3. KGP Talkie. (n.d.). [Resume (CV) Parsing using Spacy 3 | NER Training in Spacy v3]. YouTube. https://youtu.be/HJy11kOlgvk
4. Analytics Vidhya. (2021, June). Resume Screening with Natural Language Processing in Python. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/resume-screening-with-natural-language-processing-in-python/
5. Various Authors. (2023, Jul). Resume Analyser. GitHub. Retrieved from https://github.com/topics/resume-analyser