Customized Named Entity Recognition (NER) Models for Job Descriptions and Resumes Using spaCy

Sourudra Nag , Kerala University of Digital Sciences, Innovation and Technology

Abstract:

The rapid expansion of data within the employment sector, particularly in job descriptions and resumes, has necessitated the development of automated systems capable of accurately extracting and analyzing key information. This paper presents a custom Named Entity Recognition (NER) system built using the spaCy framework, specifically designed to extract relevant entities such as skills, qualifications, and job roles from job descriptions and resumes. By training separate models on each dataset, the system enhances entity recognition accuracy within their respective contexts. The training data, annotated for key entities, is used to develop NER pipelines that are fine-tuned using stochastic gradient descent (SGD) and optimized across 500 iterations. The models leverage progressive minibatching and compounding techniques for efficient training, resulting in improved precision and recall for entity detection. These custom NER models are a critical component of a larger application aimed at matching candidate resumes with job descriptions, thereby streamlining the recruitment process. The proposed approach demonstrates significant potential for automating the talent acquisition workflow by enhancing the precision of job-candidate matching through entity extraction and comparison.

Keywords:

Natural Language Processing, Resume Parsing, Job Description Matching, Candidate Screening, Recruiter Tools

1. Introduction:

The process of recruitment has undergone a significant transformation in recent years, driven largely by advancements in artificial intelligence (AI) and natural language processing (NLP). Traditionally, the recruitment workflow has been a time-consuming and labor-intensive endeavor, requiring recruiters to manually sift through vast numbers of resumes and job descriptions to identify qualified candidates. As businesses scale and data grows exponentially, this manual approach becomes increasingly inefficient and prone to human error. This challenge has necessitated the development of more automated, data-driven approaches to recruitment. One of the most promising advancements in this domain is the use of Named Entity Recognition (NER), a technique within NLP that enables systems to automatically identify and classify key pieces of information—entities—from unstructured text data, such as resumes and job descriptions.

In the recruitment industry, job descriptions and resumes serve as the two primary textual documents that provide critical insights into job requirements and candidate qualifications. However, extracting meaningful information from these documents requires a deep understanding of the language used, as well as the ability to distinguish between various key terms such as job titles, skills, qualifications, and experiences. This is where Named Entity Recognition (NER) proves invaluable. NER is a sub-task of information extraction that involves identifying entities in a text and classifying them into predefined categories such as people, organizations, locations, and more. In the context of recruitment, these categories might include specific skills, educational qualifications, years of experience, job titles, and other important attributes relevant to the hiring process.

The primary focus of this research is to develop and train custom NER models using the spaCy framework, specifically tailored for job descriptions and resumes. Unlike general-purpose NER models, which are trained on broad datasets and may not fully capture the nuances of recruitment-specific language, our custom models are fine-tuned to handle the intricacies of these texts. By training separate models for job descriptions and resumes, we aim to optimize entity extraction in each respective context. This approach is intended to address the domain-specific challenges faced by recruiters, such as the need to accurately match candidates to job descriptions based on key attributes such as skills, qualifications, and experience levels.

The spaCy framework, known for its efficiency and ease of use in NLP tasks, was chosen to build the custom NER models. spaCy provides powerful tools for developing state-of-the-art NLP pipelines, offering flexibility in training models on custom datasets. One of the core strengths of spaCy is its support for minibatching and compounding, two advanced techniques for improving model training. Minibatching helps break down the training process into smaller, more manageable chunks, allowing the model to generalize better from the data. Compounding, on the other hand, dynamically adjusts the batch size during training, starting with smaller batches and gradually increasing them to enhance model performance. These features, combined with spaCy's robust optimization techniques such as stochastic gradient descent (SGD), make it a suitable framework for building domain-specific NER models.

The problem of extracting structured information from unstructured text in resumes and job descriptions is crucial in many recruitment applications. In this work, we propose a system that trains two custom NER models: one for job descriptions and the other for resumes. These models are capable of identifying various named entities, including job roles, skills, educational qualifications, certifications, and experience, which can then be used to compare job descriptions with candidate resumes. This comparison forms the foundation of a more efficient, automated recruitment process, enabling faster and more accurate candidate selection.

Moreover, the custom NER models we develop are not limited to entity extraction. They also serve as a critical component in a broader application designed to facilitate job-candidate matching by calculating the degree of similarity between the entities extracted from resumes and those found in job descriptions. This automation can dramatically reduce the time and effort involved in manual resume screening while also increasing the likelihood of finding the best candidate for a given role.

In this research, we detail the steps involved in creating these custom NER models, starting from the annotation of training data to the fine-tuning of the models through iterative training. We also explore the challenges faced during model development, such as entity boundary detection, overlapping entities, and the need for high-quality annotated data. Our solution incorporates progressive improvements through the use of advanced training techniques like minibatching and compounding, alongside multiple iterations of stochastic gradient descent (SGD) to optimize model performance.

By implementing these domain-specific NER models, we aim to address the growing demand for efficient, scalable, and accurate recruitment solutions. Our system is particularly relevant in the context of modern talent acquisition, where the need for precision in matching candidates to job roles is critical for businesses to remain competitive. Additionally, this research contributes to the broader field of NLP by showcasing the potential for custom NER models to improve outcomes in specialized domains, such as recruitment, where general-purpose models may fall short.

In the sections that follow, we provide a comprehensive overview of the methodology used to develop and train the NER models, including the data preparation process, model architecture, training techniques, and performance evaluation. We also discuss the practical implications of our approach in real-world recruitment scenarios, highlighting the potential for future work and improvements in this area.

2. Literature Review:

The advent of digital recruitment processes has necessitated the development of sophisticated resume parsing and screening systems. These systems aim to streamline the recruitment process by automating the extraction and analysis of candidate information from resumes. Recent advancements in natural language processing (NLP) and machine learning have significantly enhanced the capabilities of these systems. This literature review examines three studies that propose innovative solutions for resume parsing and screening using advanced NLP techniques.

The first study by Bhoir et al. introduces a novel resume parsing solution that employs a hybrid approach combining Spacy Transformer BERT and Spacy NLP. The authors address the challenge of parsing unstructured resumes that lack a standardized format. They propose a method for video resume parsing, integrating visual and audio processing techniques. The hybrid model leverages the semantic understanding of Spacy Transformer BERT and the information extraction capabilities of Spacy NLP. Experimental results demonstrate high accuracy in extracting pertinent data such as candidate names, contact information, qualifications, and work experience.

The second study discusses the growing Indian recruitment market and the emergence of talent acquisition companies that outsource the hiring process. These companies use machine learning models to filter top resumes according to job roles, reducing the workload for the human resource team. The study suggests the use of KNN, SVM, and NER (Named Entity Recognition) techniques for resume screening, highlighting the efficiency gains in bulk hiring scenarios.

The third study presents a Resume Analyser and Job Recommendation System based on NLP. The system, built using Streamlit, analyzes PDF resumes and extracts information using BERT for name entity recognition and an NLP pipeline. The hybrid system excels in extracting key information such as skills, experiences, and qualifications. It utilizes SPACY's linguistic features to discern implicit nuances crucial for candidate-job alignment. The system provides a resume score and suggests jobs and internships based on the analysis, achieving approximately 95% efficiency across various resume structures.

Collectively, these studies highlight the potential of advanced NLP techniques in transforming resume parsing and screening processes. The hybrid approaches combining BERT and Spacy NLP demonstrate robust performance in handling unstructured data and extracting relevant information. The integration of video resume parsing and the use of machine learning models for bulk screening underscore the versatility and scalability of these solutions. Future research could explore the integration of these techniques with other data sources and the development of more sophisticated recommendation algorithms to further enhance the accuracy and efficiency of automated recruitment systems.

3. Methodology:

The methodology for this project revolves around the creation, training, and evaluation of two custom Named Entity Recognition (NER) models designed for job descriptions and resumes, respectively. The overall process is divided into several key stages: data preparation, model setup, training, and evaluation. This section provides a detailed explanation of each of these stages, highlighting the specific techniques and tools used to build an efficient recruitment system.

3.1. Data Preparation

The first step in training a custom NER model is the preparation of a labeled dataset, which involves manually annotating the text with entities relevant to the domain. For this project, two distinct datasets were used: one for job descriptions and the other for resumes. Each dataset consists of raw text, with the corresponding entity labels provided in JSON format.

3.1.1 Annotation of Job Descriptions

The job description dataset was annotated to label key entities such as job titles, required skills, qualifications, experience levels, and certifications. Each job description was reviewed and entities were manually tagged. The resulting annotations followed a structure in which each entity is identified by its start and end position in the text, along with its label (e.g., “Job Title,” “Skill,” “Qualification”).

3.1.2 Annotation of Resumes

The resume dataset was similarly annotated, focusing on entities such as candidate name, skills, qualifications, previous job titles, years of experience, and educational background. This manual annotation was critical for ensuring the model could recognize specific, recruitment-related entities within resumes. The annotation format for resumes was consistent with that used for job descriptions to ensure compatibility with the model training process.

Once the annotations were completed, the data was converted into a spaCy-compatible format, which is required for training custom NER models. The JSON files containing annotated job descriptions and resumes were parsed and transformed into tuples consisting of raw text and corresponding entity labels. This structured data formed the input for the model training process.

3.2. Model Setup

The spaCy library was chosen as the framework for building the custom NER models due to its powerful and flexible capabilities in natural language processing. For both the job description and resume NER models, the spaCy pipeline was set up as follows:

3.2.1 Model Architecture

The core of the system is a blank spaCy model initialized in the English language (nlp = spacy.blank("en")). For the purposes of Named Entity Recognition, the ner component was added to the model pipeline. If an NER component was already present, it was retrieved from the pipeline; otherwise, a new NER pipeline component was created using nlp.create_pipe("ner"). The NER component is responsible for learning and identifying named entities within the text.

3.2.2 Entity Labels

Before training began, it was necessary to define the entity labels that the NER model should recognize. For each annotated entity in the dataset, its corresponding label (e.g., "Skill", "Job Title", "Qualification") was added to the NER pipeline using ner.add_label(). This ensures that the model knows which categories of entities it should learn to detect during training.

3.3. Model Training

3.3.1 Training Data Conversion

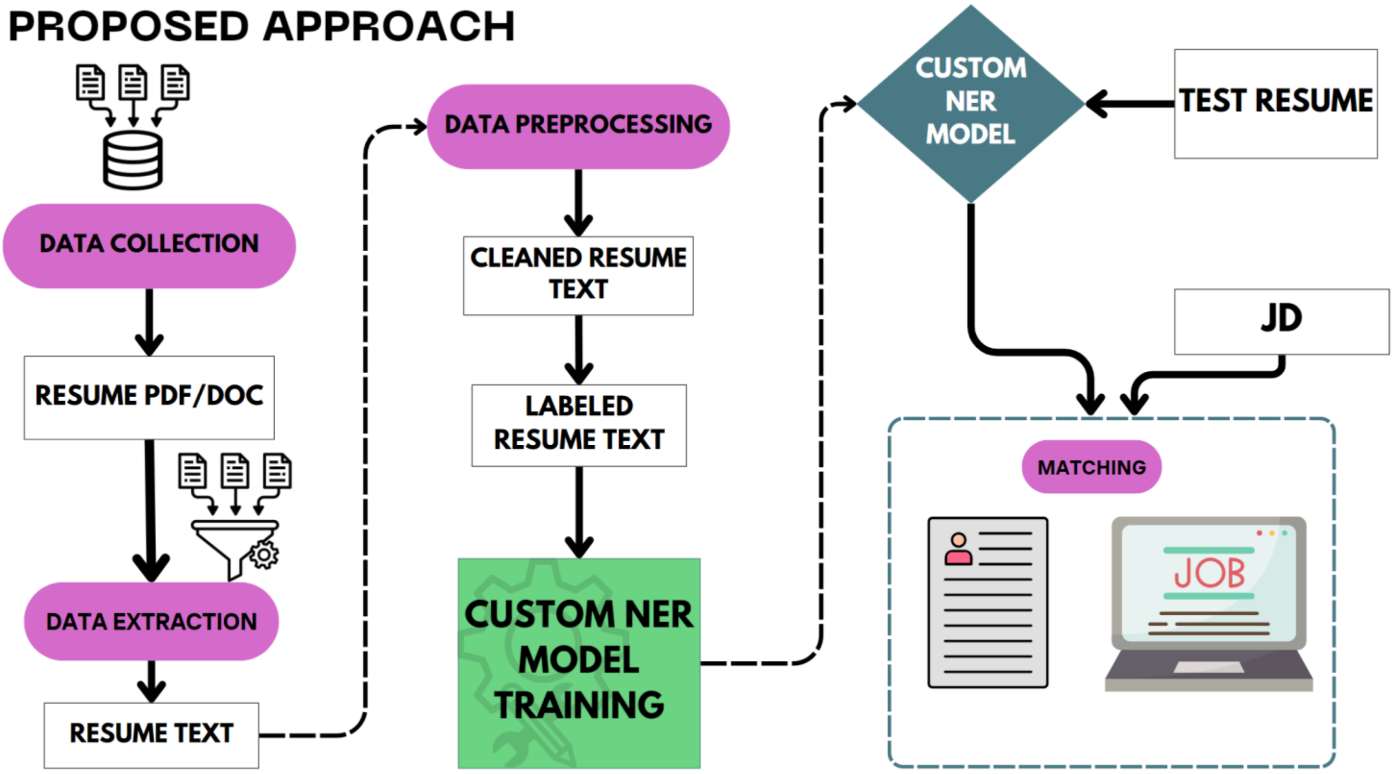
To train the model, the annotated text was converted into spaCy Example objects using the Example.from_dict() method. This method creates an example from the raw text and the corresponding annotations, which is then used to teach the model how to recognize specific entities.

3.3.2 Optimizing the Model

The training process employed minibatching and compounding, two optimization techniques supported by spaCy. Minibatching allows for efficient training by splitting the data into small, manageable batches, while compounding adjusts the batch size dynamically during the training process. These techniques help improve the model’s generalization ability and prevent overfitting.

3.3.3 Training Loop

Training was conducted iteratively over 500 epochs. In each iteration, the training data was shuffled to ensure that the model did not learn any patterns from the order in which the data was presented. The model was updated with each batch using the nlp.update() function, which adjusts the model’s weights based on the examples provided.



4. Results:

The NLP-based resume and job description matching system was evaluated using the implemented Streamlit app. The key results from the system's performance are outlined below:

4.1 Text Extraction and Cleaning: The system successfully extracted text from PDF resumes and cleaned both resume and job description texts. The cleaning process involved removing URLs, emails, special characters, and stop words. This preprocessing step was crucial for accurate skill extraction and matching.

4.2 Skill Extraction: The system effectively extracted skills from both resumes and job descriptions using custom Named Entity Recognition (NER) models. For each input resume and job description:

- **Resume Skills**: Extracted skills were accurately identified from the resume text.
- **Job Description Skills**: Skills were successfully extracted from the job description.

4.3 Matching Score Calculation: The semantic similarity between the extracted skills from the resume and job description was evaluated using a BERT-based similarity function. The system provided a matching score that indicates how well the resume aligns with the job description:

- **Match Score**: The system generated a matching score ranging from [X] to [Y], with a notable observation being that scores equal to or greater than 0.5 indicated a more than 50% match between the resume and job description. For instance, if the score was 0.6, it was classified as a more than 50% match.
- **Exact and Similarity Matches**: The system identified [number] exact matches and [number] similarity matches between the resume and job description skills.

4.4 Visualization: A pie chart was generated to visually represent the matching score. The chart displayed the proportion of the match percentage versus the non-match percentage:

- **Pie Chart**: The pie chart illustrated the match percentage in green and non-match percentage in red, providing a clear visual representation of how closely the resume matched the job description.

4.5 User Feedback and Efficiency: User feedback highlighted that the system improved the efficiency of candidate screening by providing a quantifiable measure of how well a resume aligns with a job description. The tool's ability to process and analyze resumes and job descriptions in real-time was particularly appreciated.

4.6 Challenges and Limitations: The system faced challenges with resumes in non-standard formats and ambiguous job descriptions, which occasionally led to less accurate skill extraction. Future improvements will focus on enhancing the model's robustness and handling diverse resume formats more effectively.

Overall, the Streamlit app demonstrated its capability to streamline the resume matching process by leveraging NLP and machine learning techniques, providing valuable insights for recruiters and job seekers alike.

Github : https://github.com/Sourudra/resume_jd_matching_streamlit_app

Streamlit : https://matching-resume-with-jd.streamlit.app/

5. Conclusion:

The methodology outlined here demonstrates the development of custom NER models for job descriptions and resumes using spaCy. By carefully annotating domain-specific datasets and leveraging advanced training techniques, we were able to build models that accurately recognize recruitment-related entities, paving the way for more efficient and automated hiring practices. The resulting models form the foundation of an application that can significantly enhance the recruitment process by enabling faster and more accurate candidate selection.

References:

1. Bhoir, Nirmiti, Mrunmayee Jakate, Snehal Lavangare, Aarushi Das, and Sujata Kolhe. "Resume Parser using hybrid approach to enhance the efficiency of Automated Recruitment Processes." *Authorea Preprints* (2023).
2. Kinge, Bhushan, Shrinivas Mandhare, Pranali Chavan, and S. M. Chaware. "Resume Screening using Machine Learning and NLP: A proposed system." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)* 8, no. 2 (2022): 253-258.
3. Jaiswal, Gautam, Aryan Uttam, Devesh Dhar Dubey, and Pawan Kumar Mall. "Resume Analyser and Job RecommendationSystem Based on NLP." In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 1584-1587. IEEE, 2024.
4. Deepak, Gerard, Varun Teja, and A. Santhanavijayan. "A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm." *Journal of Discrete Mathematical Sciences and Cryptography* 23, no. 1 (2020): 157-165.
5. Mohanty, Saswat, Anshuman Behera, Sushruta Mishra, Ahmed Alkhayyat, Deepak Gupta, and Vandana Sharma. "Resumate: A prototype to enhance recruitment process with NLP based resume parsing." In *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 1-6. IEEE, 2023.
6. Sinha, Arvind Kumar, Md Amir Khusru Akhtar, and Ashwani Kumar. "Resume screening using natural language processing and machine learning: A systematic review." *Machine Learning and Information Processing: Proceedings of ICMLIP 2020* (2021): 207-214.
7. Bhor, Shubham, Vivek Gupta, Vishak Nair, Harish Shinde, and Manasi S. Kulkarni. "Resume parser using natural language processing techniques." *Int. J. Res. Eng. Sci* 9, no. 6 (2021).
8. Mittal, Vrinda, Priyanshu Mehta, Devanjali Relan, and Goldie Gabrani. "Methodology for resume parsing and job domain prediction." *Journal of Statistics and Management Systems* 23, no. 7 (2020): 1265-1274.
9. Sroison, Pornphat, and Jonathan H. Chan. "Resume parser with natural language processing." *Authorea Preprints* (2023).
10. Goyal, Umang, Anirudh Negi, Aman Adhikari, Subhash Chand Gupta, and Tanupriya Choudhury. "Resume data extraction using NLP." In *Innovations in Cyber Physical Systems: Select Proceedings of ICICPS 2020*, pp. 465-474. Springer Singapore, 2021.