Resume Parsing with Named Entity Clustering Algorithm

Harshitha Badeti[1]
[1]*NLP FALL2022*

Job Application, Selections and Interviews are time consuming for both Employer and Candidates. Most of the Companies use software to scan and match Candidates' Resumes to Job Descriptions. The challenging part is the discover map right candidate to Job. This was earlier done by Recruiter's Manually scanning each Resume and map this is efficient in terms of mapping right candidate but is very time consuming in considering the pool of Resume for each Job posted. Thus, Resume Classification System (RCS) using Natural Language Processing (NLP) and Machine Learning (ML) techniques could automate this process.Automation of this process can significantly expedite and transparent the applicants' screening process with minimum human involvement. In this project, an NLP and ML-based RCS will automate the classification of the Resumes.The proposed model achieve satisfactory classification reports in terms of precision, recall and F1-score.

I. INTRODUCTION

Companies receive and process hundreds of resume for single job application.They receive application from different platforms and different formats with similar information.These resumes can be automatically retrieved and processed by a resume information extraction system which will extract useful information like name of applicant, there skills and experiences. In general Named-Entity-based,Rule-based and Learning-based methods. Usually a combination of them are used.In this paper we will look into Named-Entity using Nlp technique The identification rates of named entities such as name, e-mail and education program will in acceptable ranges or any tolerated error includes in it so because these named entities has specific or changed format.

BERT is known as a more powerful and efficient technique than the other NLP tools like RNN, CNN and LSTM, which understands inquiries better than ever before. The embedding models word2vec and GloVe have been presented to be less effective in documents recognitions. The performance also differs. In recent years, BERT has been used for sentence classification based on the suitability of job description and semantic search over the corpus. BERT will become more and more important in the staffing software

II. PROPOSED SOLUTION

A. Previous work

Existing Automated Resume scanning system are mostly built on Linux Resume parsers have been proposed to extract information from the Internet , Github and PDF , as well as from the general non-LinkedIn formats . In case of Linux, only pre-processing is required instead of the parser. Research around this has suggested Nlp key word matching and parsing can give high frequency and better processing speed when dealing with large volume of data.

B. Solution

PREPROCESSING Name Entity Recognition using Bert is proposed to do key word based search. Each entity (e.g. education information) contains a block of related information. For example, an education segment will have a number of blocks of information about educational institutions that a person attended. For example, an education information block can contain institution name, degree, major, and date information In the previous step we obtain many independent named entities .The named entity are in the block of information are need to be grouped together to do the more process on it. Named entities (chunks) are grouped according to their proximity and type. The algorithm, tries to associate related entities into a group depending on their type and how much they are close to each other.

C. Data Preprocessing

Data preprocessing is the first and foremost step of natural language processing. Data preprocessing is a technique of data mining which transforms raw data into a comprehensible format. Data from the real world is mostly inadequate, conflicting and contains innumerable errors. The method of Data preprocessing has proven to resolve such issues. Data preprocessing thus further processes the raw data. Data is made to pass through a series of steps in the time of preprocessing: **Data Cleaning:** Processes, like filling in missing values, smoothing noisy data or resolving inconsistencies, cleanses the data. **Data Integration:** Data consisting of various representations are clustered together and the clashes between the data are taken care of. **Data Transformation:** Data is distributed, assembled and theorized. Data Reduction: The objective of this step is to present a contracted model in a data warehouse. **Tokenization:** Tokenization is the task of chopping off a provided character sequence and a detailed document unit. It does away with certain characters like punctuation and the chopped units are further called tokens.

D. Comparsion with Other Models

This model is Compared against Navies based using Tf-id. When we have resumes coming in different formats Navies based methods was difficult to train based on all entities present. While Bert classified and clustered entites giving a higher accuracy.

E. Performance Evaluation

The focus in Information Retrieval research lays on text classification systems which make binary decisions for text document as either relevant or non-relevant with respect to a user's information need. Capturing the user information need is not a trivial task. We used precision, recall and F-measure metrics for performance evaluation [7]. Precision measures the number of relevant items retrieved as a percentage of the total number of items retrieved. Recall measures the number of relevant items retrieved as a percentage of the number of relevant items in collection. The F-measure is the harmonic mean of precision and recall.

The Precision, Recall and F-measure rates for segments will be predicted.

F. Future work

You can make lists with automatic numbering Now that we build Named Entity Recognition on Job seeker side if we build other Named Entity Recognition on Employer side and build a Knowledge based graph and build a recommendation system to predict the right candidate for the job.

G. Algorithm

Following are step by step procedure to solve this algorithm.

The input to the tool would be a JSON File. One doesn't have to worry about the encoding. This model extracts entities from the resum. Clean the text using our user defined cleaning function to get clean and plain text. Pass this through our trained BERT to get the result ie., the class of resume.

ACKNOWLEDGMENTS

I thank Dr. Kevin Scanell for teaching Nlp and guiding me through this project.