Automated Keyword Extraction from Job Descriptions [Draft]

Naveen Kumar Kundla
Dept. of Computer Engineering
San Jose State University
San Jose, California, U.S.A.
naveenkumar.kundla@sjsu.edu

Sivakrishna Yaganti
Dept. of Computer Engineering
San Jose State University
San Jose, California, U.S.A.
sivakrishna.yaganti@sjsu.edu

Vinit Kanani
Dept. of Computer Engineering
San Jose State University
San Jose, California, U.S.A.
vinitpankaj.kanani@sjsu.edu

***Abstract*—Applicant tracking systems (ATS) are commonly used by employers to scan resumes automatically for relevant keywords from job descriptions. However, many qualified candidates may be overlooked if their resumes lack explicit matches for keywords. This project aims to automate the extraction of key skills and qualifications from job descriptions using natural language processing. The extracted keywords can then optimize resumes to contain relevant terms that will enable candidates to pass the initial ATS screening successfully. Improving keyword matching between applicant resumes and job descriptions can help job-seekers better tailor their resumes while enabling employers to identify qualified applicants that may have been previously overlooked more accurately. Automated keyword extraction will make the initial resume screening process more efficient for both applicants and employers.**

***Index Terms*—Named Entity Recognition, natural language processing, ATS**

I. Introduction

Manually screening resumes can be a time-consuming process and may result in qualified candidates being overlooked due to the lack of specific keyword matches. This project aims to automate the extraction of key skills and qualifications from job descriptions, thereby optimizing resumes for screening by applicant tracking systems (ATS).

When you apply for a job, your resume goes through an Applicant Tracking System (ATS) that scans it for specific keywords. If your resume doesn't have the right keywords, you might not be considered for the job even if you're qualified. You can extract keywords from the job description and use them in your resume to increase your chances of getting past the ATS. This will help match your skills and experience with the job requirements and increase your chances of success.

The proposed project aims to develop a system using natural language processing techniques, specifically named entity recognition, to analyze job descriptions. The system will extract essential skills and qualifications required for the job and suggest improvements to job-seekers on how they can tailor their resumes accordingly. By optimizing the resume keywords, the chances of more candidates passing the initial ATS screening process will increase, ultimately leading to better job prospects for the candidates.

II. Literature Review

Keyword extraction from job descriptions can help job seekers optimize their resumes by identifying the most relevant skills, qualifications, and experience needed for a position. Several studies have explored natural language processing techniques to extract such keywords.

Tiwari et al. [1] present a system to extract and score entities from resumes using natural language processing and named entity recognition. It converts unstructured resume text into a structured JSON format with categorized skill sets. Harsha et al. [2] also utilize natural language processing for resume screening. It describes an end-to-end application to evaluate candidate resumes efficiently based on predefined organizational requirements. The system outputs visual representations of candidate qualification scoring to simplify review.

Previous research has focused on extracting information from and classifying/scoring applicant resume content. However, there appears to be a lack of work on optimizing resume writing by automatically extracting keywords from job descriptions. This could help job-seekers better tailor resumes and improve the accuracy of initial applicant tracking system screening.

III. Ethical considerations

A. Global Impact

The project would help job-seekers to tailor their resumes better, resulting in a more efficient job search. This would also benefit employers by enabling them to identify qualified applicants that may have been overlooked more accurately. However, the project can also be misused to bypass the initial screening process and get unqualified candidates through the door.

B. Economic Impact

This project would benefit job-seekers and employers by making the initial resume screening process more efficient. However, if the candidates misuse the project to bypass the initial screening process, it would result in higher employer costs.

C. Environmental Impact

The project would not have any significant environmental impact. The training of models would require computing resources, but the impact would be negligible.

D. Societal Impact

The project can have a significant positive impact to the well qualified candidates. However, the project can also be misused to deceive the employers.

IV. Technical Aspects

A. Dataset

1) Data Collection: This project's primary data source was greenhouse.io, a popular platform for posting and managing job openings. Automated web scraping techniques were employed to gather job descriptions, resulting in a dataset comprising over 17,000 entries.

2) Data preprocessing: From this corpus, 5000+ files related specifically to technical roles were filtered to create our final labeled dataset. This focused dataset enables training NER models optimized for extracting key details from technical job descriptions.

The dataset contains the raw text for the job title and multiple sections describing responsibilities, requirements, qualifications, etc., for each technical job listing. This serves as input for our natural language processing pipeline.

3) Data annotation: We used Label Studio, an open-source platform, for efficient and accurate dataset labeling. Leveraging Label Studio's versatile annotation capabilities, we ensured high-quality labeled data, crucial for training and validating our machine learning models.

B. NER Model

Named entity recognition (NER) is used to identify key entities and concepts in text. We experimented with pre-trained and custom fine-tuned NER models for extracting keywords from job descriptions.

1) Spacy pre-trained model: We initially tested out-of-the-box pre-trained models from the SpaCy model library [3]. The models are pre-trained on widely used NER datasets such as OntoNotes 5 and web corpus. This model supports entities like *CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME, WORK_OF_ART*, which were not relevant for our use case.

2) Custom fine-tuned model: To overcome these limitations, we annotated an initial set of 100 job descriptions to create labels for custom entities like *ROLE, SOFT_SKILL, TECH_SKILL, DEGREE, and PROG_LANG*.

We then fine-tuned the *en_core_web_md* model by training on the annotated dataset. After fine-tuning, the model was tested on the labeled job descriptions, and it achieved an F-SCORE ranging from 0.50 for the ROLE entity to 0.87 for PROG_LANG. Precision and recall also improved significantly across all entity types.

The fine-tuned model more accurately extracts entities like skills, tools, and programming languages. This enables robust extraction of keywords that can help tailor applicant resumes to each job description better.

C. Keyword extraction and Ranking

We also developed a keyword extraction and job recommendation system leveraging natural language processing (NLP) techniques. The system utilizes the trained model and the sklearn library [4] for text vectorization and similarity calculations. All the available job descriptions and the user's resume are processed to extract relevant skills and keywords. The system then computes the similarity between the skills mentioned in the job descriptions and those in the user's resume using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and cosine similarity. The resulting similarity scores are used to rank and recommend jobs to the user. Additionally, the system identifies missing keywords in the user's resume, providing valuable insights to improve the resume and increase the likelihood of matching with desired job positions. This approach enhances the efficiency of job searching and contributes to a more effective job application process.

V. Conclusion and Future Work

The results demonstrate that fine-tuning pre-trained NER models on domain-specific annotated data can significantly improve performance for extracting relevant keywords from job descriptions. The F1-scores achieved by our custom entities, such as SOFT_SKILL (0.67) and PROG_LANG (0.87), indicate that the fine-tuned model reliably extracts these critical keywords.

However, certain entity types like ROLE achieved lower F1-scores around 0.50. This implies that the model sometimes struggles to identify role keywords correctly. One potential reason is that the annotated dataset for this project comprised just 100 labeled job descriptions. More annotated examples focused on the ROLE entity could further refine the model's accuracy.

Overall, the custom fine-tuned NER model extracts keywords like skills, tools, qualifications, etc., far better than out-of-the-box pre-trained models. However, model performance varies across entity types depending on the volume of training data available. Generated keywords enable optimized resume writing, and computed job recommendation scores simplify the application process.

The dataset can be expanded by annotating more job descriptions to improve the system further, focusing on lower-performing entity categories like ROLE. Additional candidate data, such as years of experience, education level, and specialization, could be added to the job recommendation and resume optimization systems. Comprehensive profiles may improve suggestion accuracy and personalize suggested modifications even further. Larger annotated datasets and the integration of other user data could all help develop the methodologies explored in the project.

REFERENCES

[1] A. Tiwari, S. Vaghela, R. Nagar, and M. Desai, 'Applicant Tracking and Scoring System,' *International Research Journal of Engineering and Technology*, pp. 320–324, 2019.

[2] T. M. Harsha, G. S. Moukthika, D. S. Sai, M. N. R. Pravallika, S. Anamalamudi and M. Enduri, "Automated Resume Screener using Natural Language Processing(NLP)," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2022, pp. 1772-1777, doi: 10.1109/ICOEI53556.2022.9777194.

[3] English spaCy Models Documentation. https://spacy.io/models/en.

[4] Scikit-Learn, "scikit-learn: machine learning in Python — scikit-learn 0.16.1 documentation," *Scikit-learn.org*, 2019. Available: https://scikit-learn.org/.