Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 9116 - 9123

Research Article

# Resume Screening and Ranking with spaCy

**Jagadish P[1], Abhishek V[2], Anant Shukla[3], Anuj V[4], Prasanth Kumar Reddy K[5]**
Department of Computer Science, BMS Institute of Technology and Management, Bengaluru, Karnataka
[1]jaga1982@bmsit.in ,[2]1by17cs007@bmsit.in , [3]1by17cs023@bmsit.in ,
[4]1by17cs028@bmsit.in ,[5]1by17cs120@bmsit.in

## ABSTRACT

ATS is a software that provides the necessary tools for hiring companies to enable an efficient recruitment process. All recruiting companies have the irksome task of screening hundreds of resumes only to determine if the candidate is suitable for their needs. This process, done manually, is time-consuming and tedious. ATS assigns a score to each candidate based on their experience in the work field, the relevance of their experience to the recruiters and their education background. ATS takes these inputs as the primary factors in determining the eligibility. The ATS goes beyond basic spell checking and uses leading Artificial Intelligence technology to grade applications based on numerous checks that recruiters and hiring managers pay attention to. Specifically, the platform analyses application impact by evaluating the strength of word choice and also checks application style and brevity. Similarly, it also scores each of the bullet points on the application and checks for key elements such as inconsistencies, length, word choice, filler words, keywords and buzzwords.

Keywords: Application Tracking System (ATS), Artificial Intelligence (AI), Grading, Natural Language Processing (NLP), Resume Screening.

## I. INTRODUCTION

Applicant tracking systems (ATS) act as an electronic gatekeeper for an employer. The ATS parses a resume's content into categories and then scans it for specific keywords to determine if the job application should be passed along to the recruiter. Its job is to essentially weed out unqualified applicants so the recruiter can devote his or her time to evaluating the candidates who are more likely to be a match for the position. [15] describes in depth problems in hiring processes. A significant number of companies use some kind of recruitment software to speed up their hiring. This is where Artificial Intelligence-powered ATS tools come into consideration. Instead of rejecting resumes based on a few keywords, it provides applications with a score which can be used later to give applications a rank in comparison with others. The tool takes into factors such as work experience, various technical skills, marks scored throughout a candidate’s education and other numerous factors to give the application a score. This score is reflective of their skill set and experience which tells an applicant the areas they may need to improve in comparison to others. Thus, the tool is not just rejecting an applicant based on a few buzz words but instead will classify them and provide them with a score and give the recruiter a good image of the talent available to choose from. Artificial Intelligence (AI) is an add-on to the system, complementing existing best practices to provide an online recruitment solution. As the name suggests, AI enables a combination of an applicant tracking system as well as an artificial intelligence resume parsing, searching and matching engine. The result is a tool which provides accurate candidate matching vis-a-vis jobs, and ‘talent pool’ searching that helps employers streamline their hiring process.

The rest of the paper is organized as follows. In Section II, we first review some related work on ATS and resume screening, and then summarize some techniques that deal with the process involved in implementing a resume screener. Section III is an overview of our Problem Statement, here we detail the problem regarding which our research provides information. Datasets used for testing, method for training and an idea for architecture are presented in Section IV. Section V includes our general methodology and finally, we give some conclusions and insights in Section VI.

## II. RELATED WORK

### *A.* Applicant Tracking System (ATS)

ATS is a software that is needed in the data-driven times that we are in. A recent study by Glassdoor found that for every job posting a company puts out, a company receives over 250 applications [12]. Additionally, companies typically hire for more than one role at a time. That's several hundreds of resumes to go through. More often than not, many of those resumes are under-qualified [6,2,8]. Specifically, this would mean they have no relevant experience. Consider a company hiring for the position of a Java Software Developer. If the company wants someone to hit the ground running, they probably want to make sure that the candidates have at least some Java experience before they decide to hire them. ATS, helps companies filter out resumes that don't seem to have relevant experience. As we know the Indian I.T sector is the second largest candidate recruiting sector of our country. Our project addresses the Indian I.T industry but it can be extended to various other commercial sectors where intake and elimination are in bulk, like for example Governmental Jobs. To find a way to implement the project we went through the following research papers and websites to find information that plays a vital role in improving the performance of our resume screener.

### *B.* Regular Expressions (REGEX)

[13] introduces the concept of regular expression matching. Regular expression matching is used to determine if the uploaded resume includes the predefined set of keywords. The discussion is based around the implementation of regular expressions in modern Intrusion Detection Systems. However, this concept can be applied towards resume screening and developing an Application Tracking System. In [13], they present an approach for selective matching of regular expressions. Instead of serially matching all regular expressions, they compile a set of shortest patterns most frequently seen in regular expressions that allows to quickly filter out events that do not match any of the IDS signatures. They developed a method to optimize the final set of patterns used for selective matching to reduce the amount of redundancy among patterns while maintaining a complete coverage of the IDS signatures set. The pattern recognition concept involved in IDS signatures can be applied to our Application Tracking System. This technique resulted in 26% increase in performance on average.

### *C.* Named Entity Recognition (NER)

We receive unstructured data while performing resume screening. This data cannot be handled by standard data types since a considerable amount of time to segregate data would be wasted. To avoid this, we use a concept we came across called Name Entity Recognition. NER is an important subtask in natural language processing. Various NER systems have been developed in the last decade as detailed in [2]. It also discusses the Hidden Markov Model (HMM) as well as LSTM-CRF model. A bi-LSTM-CNN model is discussed as well. They may target different domains, employ different methodologies, work on different languages, detect different types of entities, and support different inputs and output formats. [4] discusses the essential details about spaCy. spaCy's NER tool extracts named entities in nineteen categories: technical skills, persons, nationalities or religious groups, facilities, organizations, geopolitical entities, locations, products, events, works of art, law documents, languages, dates, times, percentages, money, quantities, ordinals, and cardinals. From a technical perspective, spaCy utilizes a set of well-established entity recognition models that are based on statistical learning methods to identify named entities in texts.

### *D.* Ranking

After extracting data and sorting data using Regular Expressions and Name Entity Recognition, we must rank the obtained dataset according to the discretion of the employer. We discovered a method to do this in python. It can be done using a software library in python called Pandas. We referred to the pandas dataframe website to obtain this information. In pandas we can create a 2-Dimensional data structure called dataframes. This dataframe will contain the ranking of the screened resumes. A rank function in pandas provides us various tools to successfully implement the ranking system.

The tools are:

•axis: Index to direct ranking.

•method: How to rank the group of records that   have the same value (i.e., ties)

•numeric only: For DataFrame objects, rank only

### *E.* Skill Extraction

Extracting explicitly mentioned skills is quite easy while performing resume screening, obtaining the details that are not explicitly mentioned is a very hard task. [1],[11] introduces the concept of extracting implicit skills-the skills which are not explicitly mentioned in a Job Description but may be implicit in the context of geography, industry or role. Results for matching resumes and Job Descriptions demonstrate that the proposed approach gives an improvement of 29.4% when compared to the performance of a baseline method that uses only explicit skills. [1] also provides a detailed explanation on how to perform implicit skill extraction. [9] details an ontology-based approach (ORP) which can also be employed as a supplemental aid.[14] provide valuable insight to enable extraction of information from the resulting semi structured data.

## III. PROBLEM STATEMENT

ATS is a recruiter's gain in terms of speed of their hiring processes. The sheer amount of time and work that goes into scanning each one resume manually is cumbersome and is not ideal. ATS goes beyond basic spell checking and uses leading AI technology to grade applications based on numerous checks that recruiters pay attention to.

ATS assigns a score to each candidate based on their experience in the work field, the relevance of their experience to the recruiters and their education background. Where these scores help the organization to make out the best possible candidates list according to their given constraints and requirement for that particular vacancy, implementation of AI-powered tools that allow no candidate to be overlooked and provide an equal opportunity for all based on their skill set must be the primary goal.

## IV. DATASET AND MODEL ARCHITECTURE

### A. Test Dataset

Test Dataset includes a set of documents and a set of PDFs. Each document contains a resume which includes the work experience and other details the candidate wishes to share with the recruiter.

We conduct the test by stating the parameters for recruitment and based on the final ranking of the resumes we will be capable of analysing the results. For the test to be successful, the ranks allotted to resumes must match the requirements of the recruiter. If the ranks are a mismatch, further improvement in learning is necessary.

### B. Training Method

SpaCy uses a deep learning formula for implementing NLP models, summarised as “embed, encode, attend, predict”.

In spaCy’s approach text is inserted in the model in the form of unique numerical values (ID) for every input that can represent a token of a corpus or a class of the NLP task (part of speech tag, named entity class). At the embedding stage, features such as the prefix, the suffix, the shape and the lowercase form of a word are used for the extraction of hashed values that reflect word similarities. In [3], the authors explain in great detail why spaCy is an excellent training method for language processing.

NLP is used to analyse and extract information in a resume.

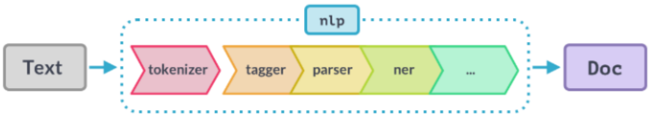

**Fig 1: spaCy working model**

This representation showcases the NLP approach by spaCy. However, in this particular case of ATS; we use spaCy to screen a document to provide us with required text. This text can be matched with constraints using the KMP string matching algorithm.

### C. Architecture

Fig 2 shows the proposed architecture. The ranking of the resumes is done in two phases. Phase 1 deals with assigning and id and scores to the person's personal experience, university scores etc. Phase 2 deals with the allocating a score and experience depending on their skill. A ranking score determines the order of these resumes and is arranged into 3 subsections, the best suited resume in section 3, next in section 2 and so on.

## V. METHODOLOGY

The resumes must be analysed to extract the predefined set of properties. These properties are subject to changes by the recruiter and it acts as the search filter and enables the recruiters to find suitable employees. Skills are extracted from the resumes provided using the predefined technologies such regular expressions, and spaCy. Various fields such as work experience, employees experience in the field in terms of years are used. This is used to determine rankings. Rankings are intra section and this is based on skill experience and skill score that can be determined using the set parameters. Defining various criteria for ranking is very necessary to ensure successful implementation of said project. Upon completion of ranking within employees. The various companies and universities the employees have worked in are assigned their own ranking specified here as inter-ranking based on which the companies and universities are analysed. An employee’s total approval rating will be inclusive of inter- section and intra-section ranking.

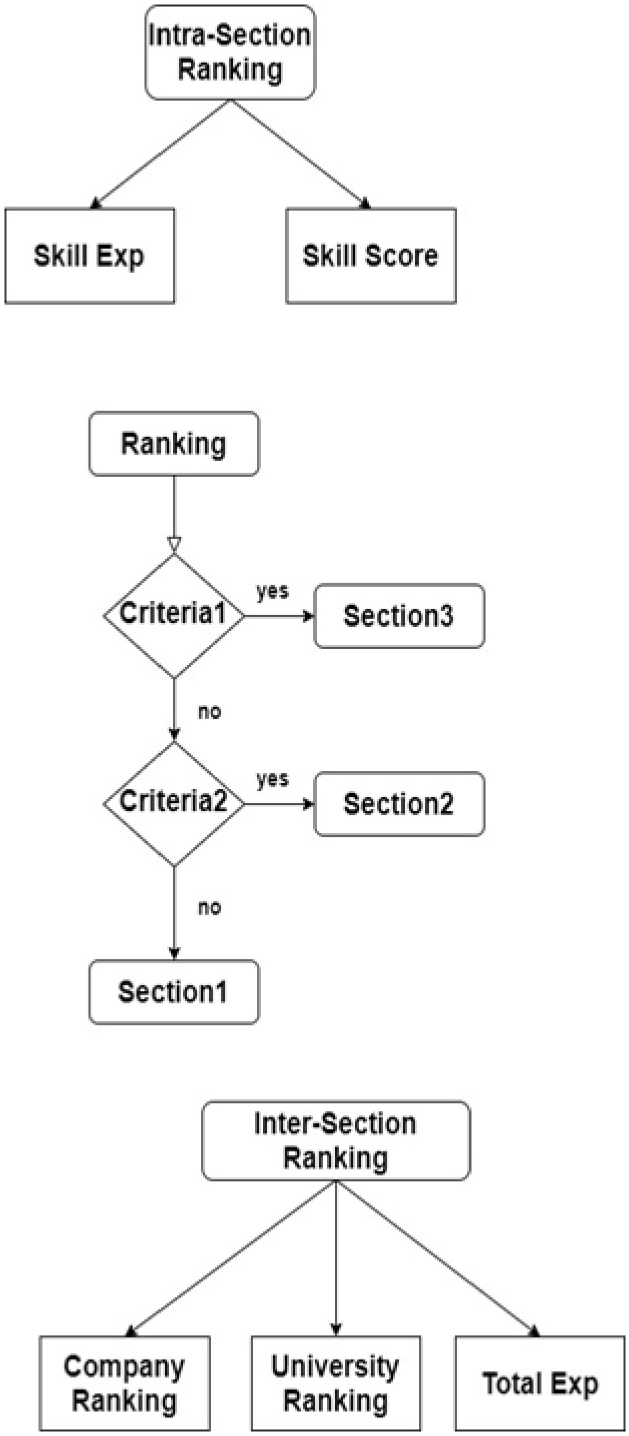

**Fig 2: The Proposed Architecture.**

## VI. CONCLUSIONS

NLP and spaCy has solved multiple problems and have various applications. These applications include name entity recognition, and data analysis using Artificial Intelligence. In this paper, we propose an idea and briefly describe a methodology for implementing an ATS using Regular Expressions, NER and String-matching algorithms. Although a significant amount of work has been done in this field, implementing said ideas using AI provides a significant improvement in ranking, consequently helps in refining existing techniques.

## ACKNOWLEDGEMENT

This work has been done by a 4-student group supported by a professor of BMS Institute of Technology.

## REFERENCES

[1] Gugnani, Akshay & Misra, Hemant, “Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation”, Conference: 2020 AAAI - Innovative Applications of Artificial Intelligence (IAAI).

[2] Y. Su, J. Zhang and J. Lu, "The Resume Corpus: A Large Dataset for Research in Information Extraction Systems," 2019 15th International Conference on Computational Intelligence and Security (CIS), Macao, Macao, 2019.

[3] E. Partalidou, E. Spyromitros-Xioufis, S. Doropoulos, S. Vologiannidis and K. I. Diamantaras, "Design and implementation of an open-source Greek POS Tagger and Entity Recognizer using spaCy," 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Thessaloniki, Greece, 2019, pp. 337-341.

[4] C. H. Ayishathahira, C. Sreejith and C. Raseek, "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing," 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, 2018, pp. 388-393, doi: 10.1109/CETIC4.2018.8530883.

[5] Y. Luo, H. Zhang, Y. Wang, Y. Wen and X. Zhang, "ResumeNet: A Learning-Based Framework for Automatic Resume Quality Assessment," 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 2018, pp. 307-316, doi: 10.1109/ICDM.2018.00046.

[6] V. Lai et al., "CareerMapper: An automated resume evaluation tool," 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 2016, pp. 4005-4007, doi: 10.1109/BigData.2016.7841091.

[7] Jiang et al., “Evaluating and Combining Name Entity Recognition Systems”, 2016, Proceedings of the Sixth Named Entity Workshop, pp. 21-27, doi: 10.18653/v1/W16-2703.

[8] S. Bremner and B. Phung, "Learning from the Experts: An Analysis of Résumé Writers' Self-Presentation on LinkedIn," in IEEE Transactions on Professional Communication, vol. 58, no. 4, pp. 367-380, Dec. 2015, doi: 10.1109/TPC.2016.2519319.

[9] D. Çelik et al., "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs," 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Japan, 2013, pp. 333-338, doi: 10.1109/COMPSACW.2013.60.

[10] Stakhanova, Natalia & Ren, Hanli & Ghorbani, Ali, “Selective Regular Expression Matching”, Conference: Information Security - 13th International Conference, ISC 2010, Boca Raton, FL, USA, October 25-28, 2010, doi: 10.1007/978-3-642-18178-8_20.

[11] Z. Chuang, W. Ming, L. C. Guang, X. Bo and L. Zhi-qing, "Resume Parser: Semi-structured Chinese Document Analysis," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, 2009, pp. 12-16, doi: 10.1109/CSIE.2009.562.

[12] Glassdoor.com “50 hr recruiting stats that make me think” [Online]. Available: https://www.glassdoor.com/employers/blog/50-hr-recruiting-stats-make-think/.

[13] Jiang, Ridong & Banchs, Rafael & Li, Haizhou, “Evaluating and Combining Name Entity Recognition Systems”, Conference: Proceedings of the Sixth Named Entity Workshop, doi: 10.18653/v1/W16-2703.

[14] V. K. Ravindranath, D. Deshpande, K. Venkata Vijay Girish, D. Patel, N. Jambhekar and V. Singh, "Inferring Structure and Meaning of Semi-Structured Documents by using a Gibbs

Sampling Based Approach," 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 2019, pp. 169-174, doi: 10.1109/ICDARW.2019.40100.

[15] Rozario, Sophia & Venkatraman, Sitalakshmi & Abbas, Adil. “Challenges in Recruitment and Selection Process: An Empirical Study”. doi: 10.3390/challe10020035.