Developing an Effective Question Answering System for Job: Bridging the Gap Between Job Seekers and Employers

Sugam Sharma
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
0009-0008-8526-9119

Apurva Patil
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
0009-0000-8996-5121

Mayur Gaikwad
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
mayur.gaikwad.phd2019@sitpune.edu.in

Shruti Patil
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
shruti.patil@sitpune.edu.in

Ketan Kotecha
Department of Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Pune, India
director@sitpune.edu.in

***Abstract*—In the digital age, the abundance of job-related information available online can overwhelm job seekers, necessitating a streamlined solution. This research paper addresses the development of a question-answering system tailored to job descriptions. The system's primary objective is to facilitate efficient and accurate information retrieval for job seekers, improving the matching of candidates with employers. By harnessing natural language understanding and information extraction, this system simplifies the job search process and enhances the user experience, ultimately contributing to a more efficient and less frustrating job-seeking and hiring process.**

***Index Terms*—**

I. Introduction

In today's digital age, the internet is brimming with websites that offer a plethora of information related to job opportunities. Platforms like Glassdoor, Unstop, Job Search India, and Indeed are just a few examples of the many websites that cater to job seekers. However, the abundance of choices can often leave individuals feeling overwhelmed, unsure of where to turn for reliable and comprehensive job-related information.

Now, consider the prospect of having a single, centralized platform that serves as a holistic resource for all your job-related inquiries. This proposed platform goes beyond the conventional role of merely facilitating job applications. It aims to become a comprehensive repository of information, offering insights into the intricacies of the job market. This includes not only details on how to apply for a job but also valuable information such as the specific qualifications and skills demanded by companies, the salary ranges associated with various job positions, and the broader job market trends.

In this scenario, a question-answering bot takes center stage, acting as an indispensable tool for information retrieval. Rather than embarking on extensive and often convoluted Google searches, users can now rely on this chatbot as a dedicated and efficient source for job-related information. It simplifies the process, condensing the wealth of information into straightforward, easy-to-understand responses, creating a user-friendly experience in the ever-evolving landscape of job seeking.

This problem statement emphasizes the importance of making a system that helps job seekers easily find the right information in job descriptions. This will make applying for jobs easier and help match candidates with the right jobs. The research paper will look at different challenges like understanding natural language, extracting information, and creating a user-friendly system for job seekers and employers.

II. Goal of the Paper

The motivation behind researching and developing question-answering system for job descriptions is multifaceted. In today's information-rich digital landscape, the sheer volume of job listings can overwhelm job seekers, making it challenging to identify the most relevant opportunities. Thus, a system that streamlines this process is sorely needed. Such a system offers a time-efficient solution, saving job seekers the arduous task of combing through countless job descriptions by quickly providing them with the information they need. Furthermore, it can significantly enhance the matching of candidates with employers by increasing the accuracy of job-seeker and job-post alignment, reducing missed opportunities on both sides. The motivation also extends to improving user experience, as a user-friendly interface and a system that understands natural

language can make the job search and hiring processes more efficient and less frustrating.

III. Literature Review

In the realm of question-answering systems, several notable systems have been developed to tackle a range of challenges. These systems vary in their capabilities, applications, and the hurdles they face. This literature review will provide an overview of these systems, their key features, and the associated challenges.

Green et al. (1961) introduced the BASEBALL system, which specializes in answering questions related to dates, locations, and American baseball games. The principal challenge faced by this system is the manual creation of its knowledge base. [1]

Woods (1973) presented the sLUNAR system, designed to support geological analysis of rocks from the Apollo mission. Similar to BASEBALL, sLUNAR relies on manual knowledge base creation, which can be resource-intensive. [2]

Mollá and Vicedo (2007) described the Modern QAS systems. These systems excel in answering natural language questions and encompass diverse dimensions such as the source of answers, the domain of application, and question analysis methods. However, they grapple with challenges associated with data expansion, domain specialization, and the increasing complexity of questions. [3]

Dan Mollá's (2002) paper, "Question Answering in Large Technical Corpora," explores the challenges and methodologies of creating question-answering systems for extensive technical datasets. The study focuses on techniques akin to Named Entity Recognition and relation extraction, emphasizing the importance of precision in answer extraction. It addresses the complexities of domain-specific language and the need for accuracy, scalability, and broad topic coverage. Mollá's work is a valuable contribution to improving question answering in specialized domains. [4]

Rinaldi et al. (2004) introduced Biomedical QA systems, including the Extra Ans system, which incorporates Minimal Logical Form for genomics applications. The primary challenge here is aligning semantic representations with the knowledge base, ensuring accuracy in biomedical contexts. [5]

IV. Methodology

A. Data Collection

Data collected from various websites using the Selenium web automation tool has been compiled into a single CSV file. The dataset includes academic and technical attributes:

- Company: Company's name.
- Education: Required educational qualifications.
- Experience: Minimum experience needed.
- Industry: Company's sector.
- Job Description: Detailed job role description.
- Job ID: Unique identifier for each posting.
- Job Location: Geographical work location.
- Job Title: Official job title.
- Number of Positions: Vacancies available.
- Pay Rate: Compensation details.
- Site Name: Source website.
- Skills: Required skills.
- Unique ID: Unique posting identifier.

This dataset is valuable for academic and technical analyses, including job market trends, skill demand, and educational requirements. It can support research, machine learning, and data-driven decision-making in various domains.

B. Data Preprocessing

In the data preprocessing phase, we follow a systematic approach to handle missing values and derive a new column for job roles from the existing job description column. Specifically, our methodology is as follows:

- Handling Missing Values: We begin by addressing columns in our dataset with a high percentage of missing values, defined as exceeding 70%. To maintain data quality and avoid any potential distortions, these columns are removed from the dataset.
- Imputing Missing Values: For columns where the proportion of missing values is less than 70% we employ a standardized approach to fill these gaps. Specifically, missing values are replaced with the placeholder term "unknown." This ensures uniformity and enables us to work with complete data for subsequent analysis.
- Creating the 'job_role' Column: We introduce a new column in our dataset called 'job_role,' which encapsulates information extracted from the 'jobdescription' column. The 'jobdescription' column contains various flags representing different features of job postings. Notably, the 'Education' feature is signified using two distinct flags: 'Education Qualification-' and 'Education-.'
- Standardizing the 'Education' Flag: To ensure consistency, we transform occurrences of 'Education Qualification-' to 'Education-' within the 'jobdescription' column. This alignment allows for streamlined analysis and simplifies the extraction of job roles.
- Job Role Extraction: The 'job_role' column is populated by extracting relevant content from the 'jobdescription' column. We identify the start of a job role by detecting the 'Role:' flag and determine its conclusion at the 'Education-' flag. This process enables us to isolate and store job roles separately.
- Handling Garbage Values: We recognize that the 'jobdescription' column may contain extraneous characters and symbols such as '\s', '\xa0', '_', '—', and '.'. These elements are removed from the extracted job roles, ensuring that the 'job_role' column contains clean and standardized job role descriptions.

By following these precise steps, we enhance the quality of our dataset, enabling more rigorous and meaningful academic and technical analyses of job roles and associated information.

C. Data Formatting

Data needs to be formatted to make it suitable for the transformer model. First, every column of the data has been

joined together with added flags in front of every column. For e.g. jobtitle is joined with company column by adding 'JT_' and 'CO_' infront of jobtitle and company column respectively so that it becomes easy later to find index of the different sections. This combined data will act as a context data. Now we create a index dataframe which will store different indexes of the starting of different columns of context data based on flags we have made. For creating questions for the context we have used 5 different types of questions. Each context will follow same 5 types of question patterns. Question patterns includes

- "What is the job provided by X company?"
- "What is the salary provided by X company for Y post?"
- "What are the skills required for Y post in X company"
- "What are experiences required for Y post in X company
- "What is role for Y post in X company"

The X and Y are replaced by the values in each context fror context data with the help of index dataframe. The data i formatted into json as required by simple transformer library Finally json formatted data is splited into 80:10:10 ratio int training, testing and validation set respectively.

D. Model Training

We utilize the SimpleTransformer library to train a Questio Answering model. With learning rate of 5e-6, the transforme model is trained till 2 epochs as shown in 1.

E. Prediction

We aim to make a question-answer system onto which user asks question and context is dynamically choosen from the existing context dataset. In oder to implement that, context dataset is converted into vectors using gensim glove-wiki-gigaword-200. For making context vectors, each context is first preprocessed by lower casing, removing punctuations, accent words and stopwords as showin in 2. Then the context data is transformed using glove vectorizer. Next, we pre-process the question using same pipeline as we used to process context and transform the question into vector using the same glove model. We then compute the cosine distance between the question vector and every context vector. The context vector with the highest cosine similarity is selected as the actual context to the given question. The actual context and question is formatted into json format and passed into the pre-trained model for prediction.

V. Results and Discussion

Our question-answering bot for job descriptions has resulted in some interesting outcomes and enlightening discussions: Increased Efficiency: Our approach has actually revolutionized the job hunting process. It is no longer necessary to spend hours searching through job ads. Instead, our bot swiftly collects critical information from job descriptions, such as credentials, abilities, and pay information, making the job search more efficient. Better Matching: Our method improves the match between applicants and businesses by accurately linking job seekers with job posts that match their skills and

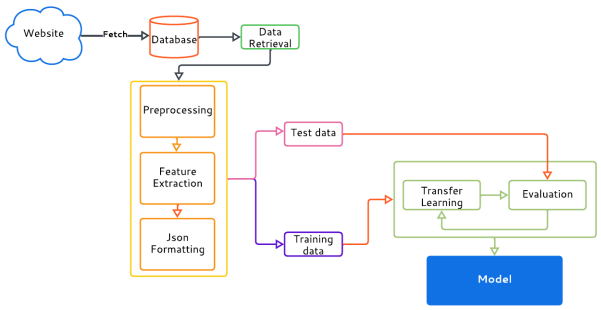

Fig. 1. Training Workflow

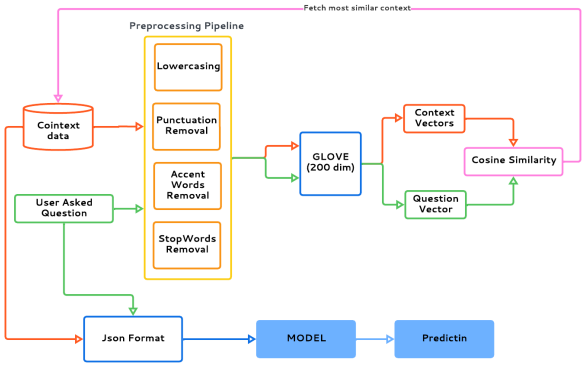

Fig. 2. Testing Workflow

credentials. This reduces the number of missed chances and raises the overall quality of job applications. User-Friendly Experience: We built our chatbot to understand normal language and respond in simple, easy-to-understand ways. Say goodbye to the perplexity of complicated job search websites. Finally we have used Cohen Kappa to evaluate our model. We achieved κ of 0.83. The value of 0.83 shows near perfect agreement on the performance of the model as shown in I.

VI. Limitation

The data is limited to a two-year timeframe, meaning our question-answering bot cannot process job descriptions from before 2021.Currently, we have created a dataset using five sets of questions for each context. However, all the questions within the dataset follow the same pattern, which can lead to overfitting of the model. Creating the dataset is a time-consuming task, so we have implemented a more efficient approach by using similar questions.

VII. Conclusion

The development of a question-answering system tailored for job descriptions holds immense promise for the future, offering several compelling advantages. Firstly, in an ever-evolving job market, this system will be invaluable in helping job seekers swiftly identify the most relevant opportunities,

TABLE I
TWO-WAY FREQUENCY TABLE FOR OBSERVATION

		Person B		
		Correct	Not Correct	Total
Person A	Correct	600	20	620
	Not Correct	60	320	380
Total		660	340	1000

thereby reducing job search time and increasing the likelihood of finding the right fit. Additionally, employers stand to benefit by streamlining their recruitment processes, saving time and resources while improving candidate-employer alignment.

REFERENCES

[1] J.R. Green, et al., "BASEBALL: Answered questions about dates, locations, and American baseball games," in Manual knowledge base creation, 1961.

[2] R.L. Woods, "sLUNAR: Supported geological analysis of rocks from Apollo mission," in Manual knowledge base creation, 1973.

[3] D. Mollá and J.L. Vicedo, "Modern QAS: Answering natural language questions is a major challenge due to data growth. Diversified dimensions: answer source, domain of application, and question analysis methods," in Coping with data expansion, domain specialization, and question complexity, 2007.

[4] D. Mollá, "Question Answering in Large Technical Corpora," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 222-229.

[5] Rinaldi et al. (2004), "Biomedical QA: Extra Ans system and Minimal Logical Form for genomics application," in Alignment of semantic representations with the knowledge base, 2004.