Literature Review of Skill Extraction Studies

By Adeluyi Aderogba

The overall economy of a country is largely dependent on the productivity of its citizens. With the vast amount of businesses, the recruitment of quality staff is essential for the business' attempt of reaching maximum productivity while simultaneously maximizing their revenues and profits. Therefore, it is important to not only recruit the right staff, but have a quality recruitment process. Being in a digital world, the use of online platforms and systems is becoming more and more important. To improve recruiting methods, there needs to be an adequate match between the Job Description (JD), and the skills provided by the job seeker. The Job seeker's skills are presented through the form of a Curriculum Vitae (CV). The essence of this report is to not only extract essential skills relative to the JD, but also develop a system that will match the JD and CV accurately. This system will be of huge benefit to firms and their heads of recruitment. As part of my report, I researched past studies in skill extraction, the following sections will provide summaries, comparisons, and the strengths and weaknesses of these studies.

Section 1: Major Challenges

Through the use of Natural Language Processing (NLP) techniques, I intend to extract the necessary skills that are mentioned in both the JDs and CVs. As part of the task to match job descriptions (JDs) to the Curriculum Vitaes (CVs) of individuals I envision that there will be significant challenges. The following will highlight challenges that I will come across in my attempt to carry out the skill extraction.

Structure of JDs and CVs

One of the main challenges that I anticipate to run into is the structure of Job Descriptions and Curriculum Vitaes. Having researched JDs for different roles, it is clear that there is no universal format for the way in which these JDs are set up. So in order to combat this challenge, I intend on extracting the keywords from the job description that represent skills. Therefore, when a JD for a financial analyst includes "Adequate at Financial Modelling ", I will specifically look for the key skill words or phrases that apply to the skill such as "Financial Modelling". This will allow our algorithm to successfully extract the essential skills and match it to the CVs on offer from the candidate.

First of all, the structure of CVs differentiate based on the country of the candidate. Depending on where I retrieve the CV dataset from, there will be a variety in the format of the CVs. In order to do this, certain candidates are creative in the way they put in all the skills they want to portray to the recruiter. Knowing this, I will use the same strategy used for the JDs. We specifically search for keywords/phrases that will be extracted and defined as a skill.

Typing Errors and Different Meaning/Spelling of Terms

Another challenge I will face is the fact that certain skills will have both multiple types of spellings and multiple meanings. Having conducted research on past approaches to the skill

identification tasks, the multi meaning and spellings of key skills is a recurring challenge. An example of this is the use of certain programming languages. The programming language, C++ can be written as “C++” however certain candidates might choose to spell it as “c plusplus” or “Cpp”. I also have to take into consideration that there will be candidates that misspell certain skills or have typing errors in their. For example “Python” could be spelt as “Pyton” or “Phyhton”. Therefore the skill extracting model I will construct will have to account for these potential typing errors or different spelling of terms.

Section 2: Review of Related Works

As part of our task to construct a skill extracting model, the research of existing studies will be a huge asset as I proceed with the task. There are various studies in this field that thoroughly explain the procedures in skill extraction, maneuvering through the challenges as well as providing its application in the recruitment of staff. A summary of these studies will be provided in this section, highlighting the purpose of the papers, challenges the authors came across, and their methodology.

Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation by Akshay Gugnani, and Hemant Misra.

This paper by Gugnani and Misra does a good job in their quest to match Curriculum Vitaes (CVs) with Job Descriptions(JDs). The mission of these two authors was to create a job recommender system that will provide candidates with the ability to receive a job recommendation based on the existing skills they possess and the requirements of the industry they apply to.

Similarly to the challenges mentioned in section 1, Gugnani and Misra encountered significant challenges at the start of their research. The challenges encountered before the start of the study for these authors included: CVs and JDs are typically not written the way well-formatted articles are written in newspapers etc; CVs may contain tables and other formatting features to make them look attractive, but this makes it difficult to obtain relevant information from them; Matching skill keywords between CVs and JDs may not yield good results because of the complex link between the skills; The JDs may be too descriptive or too simplistic, and may not uncover the essence of the offered positions and roles.

The authors highlighted the methodologies adopted to overcome the challenges they encountered. Below is their methodological process:

1) **Data collection:** Prior to implementing their strategy, Gugnani and Misra had to get accurate data to be able to test certain aspects of their system. Web scraping is a technique that was used to create a dataset. They were able to curate a dataset that contained 1.1 million mined Job Descriptions from different industries including technology, education, finance etc. To find CVs that will match their mass amount of

JDs they have, they borrowed a CV dataset which consisted of 1314 different CVs. Finally as they intend on using Natural Language Processing, they retrieved a skill dictionary which provided them with a comprehensive list of skills that they are able to use when extracting the key skills from JDs and CVs.

2) **Natural Language Process (NLP):** In order to perform the skill extractions the use of NLP was essential. Within the skill extraction module, the following NLP models were used:
    i) **Named Entity Recognition (NER):-** The NER is a subtask of information extraction that seeks to locate and classify terms occurring in natural text into predefined categories such as names of persons, organizations, locations, etc. The keywords that were identified as a result of the NER were used to validate and identify noun-phrases as skills or technological terms.
    ii) **Part of Speech (PoS) tagger:-** This model is used for the identification of nouns, verbs, adjectives, adverbs, etc. However within this model there involved some manual work to identify relevant skills. The use of domain experts helped the authors with their identification of skills. These experts read various JDs and identified which terms or phrases they identified as skills. This manual annotation was then processed through POS tagger in order to identify the nouns, verbs, adverbs, etc.
    iii) **Word2Vec (W2V)**:- A group of related models that are used to produce word embeddings. The W2V model was essential in identifying and learning new skill phrases. This model also allowed the authors to carry out their assumptions that a skill-term could be a set of words or phrases. This was so as the W2V model tokenises using white spaces.
3) **Matching JD and CV :-** Prior to matching the JD and CVs, the authors included the idea of implicit skills. Implicit skills were defined by them as a skill which has not been directly or explicitly mentioned in a JD but may be relevant for the job role. Following the acknowledgement that there is a discrepancy in the skills required in job descriptions which will lead to the poor matching of JDs and CVs, the authors decided to compute an affinity score to remove the variance when matching both CVs and JDs. The affinity score measures how suitable the recommendation is for the given candidate and JDs, and ranges from [0,1] with 0 being a poor match and 1 being a good match. The results of applying this system when trying to match JDs and CVs for 25 applicants, shows that there is a 0.88 accuracy for implicit skills for the first recommendation. A 0.96 accuracy at the third recommendation, and a 1 accuracy at the fifth recommendation. When incorporating the system to a larger database there is a 0.85 accuracy for the first recommendation, 0.95 accuracy for the third recommendation, and then a 0.98 accuracy for the fifth recommendation.

The system of matching JD and CV works for the authors as they were able to achieve high accuracy numbers based on the datasets used. It still remains to be seen if the models used can be applied and be accurate for more diverse and larger datasets. Some of the

improvements suggested by the authors included: the use of a more diverse dataset, a stronger evaluation of the system by including more CVs and leveraging additional techniques for skill identification and extraction. Also they suggested the use of skill graph for inferring professional growth of a user and leveraging that for better recommendations.

LinkedIn Skills: Large-Scale Topic Extraction and Inference by Bastian et. al

This study provides the methodological process when attempting to develop a large-scale skill extraction pipeline, therefore implementing an inference and recommender system for skills.

As with all skill extraction studies, there were challenges prior to the start of the study. These challenges included: Hundreds of millions of profiles contained millions of keywords that might or not might be skills; Many of the potential skills were duplicates. For example “Java Programming” and “Java development” both mean the ability to code Java, however that is being expressed in multiple ways; Phrases have different meanings, depending on the context. “Organ” could mean a musical instrument to a musician or a part of the body to a physician; Typing errors of skills. For example “Government liaison” and “Government liason”.

The methodology section is divided into two parts by the authors. One part being Skill and Expertise, while the other is Skills Inference and Recommendation.

- **Skill and Expertise:-** A folksonomy was created as part of the data extraction process. This was done so as to categorise the skills collected. For example “ability to use excel”, “expert at financial modelling” are skills that are essential for the role of a Financial Analyst, therefore it will be under the category “Finance”. The following were the three phases conducted when creating the folksonomy:
    A. **Entity extraction:-** This phase involved the extraction of phrases that could be considered skills. These phrases were taken from profile speciality sections which contained comma-separated lists of topics.
    B. **Clustering to provide context:-** this section involved the disambiguation of phrases that had multiple meanings. This included the generation of a set of related phrases for candidate skill phrases. These phrases are phrases that appear often with another skill phrase on profiles. Therefore the words that are considered ambiguous will have related phrases that will provide the needed context to understanding their true meanings. Clustering was applied to distinguish the distinct senses for each phrase.

C. **Crowdsourcing:-** This phase was done to combat the challenge of typing errors and word duplications found in the CVs and JDs. The authors used wikipedia and Amazon's Mechanical Turk to help create crowdsourcing solutions.

A Graph-Based Approach to Skill Extraction from Text by Kivimaki et al.

The authors of this study constructed a skill extraction system that identified skills from datasets through the use of input texts and wikipedia. In order to develop the skill extraction system, the authors collected data from Wikipedia and Linkedin. They used the hyperlinks and encyclopedia aspects of wikipedia in order to mine the data. The skills to look out for were collected from Linkedin. This was done so as to be able to validate which word/phrase is considered a skill.

The others encountered similar challenges that previous authors had encountered when trying to extract skills. The challenge in this instance was that some skills were identified as hubs of the Wikipedia graph, constantly retrieved, regardless of what the input was. As this challenge was as a result of their approach, the authors had to adopt a spread activated algorithm to overcome their challenge.

The skill extraction system, Elisit, integrated both the text2wiki, and wiki2skill models. The text2wiki module's goal is to retrieve Wikipedia articles that are relevant to the input text. The wiki2skill uses the input text that is used in the text2wiki module, returning a vector of final activations of all the nodes of the network and a vector containing the activations of only the nodes' corresponding skills.

Section 3: Comparison and Contrast of Studies

Having read and reviewed some existing studies, this section will compare and contrast aspects of these reports that we feel will be essential to our report.

Data Collection

The first section of a data analysis report is the collection of data. Due to the fact that most of these reports are done by secondary parties, the data collection is gotten from secondary sources. Gugnani and Misra adopt a data collection technique of web scraping. In doing this they were able to retrieve 1.1m JDs from different industries and 1314 CVs that will be looked at to match the JDs. Web Scraping is the extraction of data through websites, and is a technique we see with other reports such as 'LinkedIn Skills: Large-Scale Topic Extraction and Inference. Bastian et. al' and 'A Graph-Based Approach to Skill Extraction from Text'.

Acknowledging the fact that going through these reports serves as a literature review for my report, the aim of most reports is to build a system that will help with the extraction of skills. However the method in which these systems are created is diverse. One constant in the

extraction of these skills is the use of Natural Language Processing (NLP). This enabled the authors to be able to name, categorise, and then identify the skills. However, to know that a noun, verb, adjective is a skill word/phrase, there had to be a dictionary that defined them. Gugnani and Misra created a skill dictionary in order to achieve this. However, Bastian et al. and Kivimaki et al. both used the websites to determine what a skill was. Both studies used Linkedin's database as a source, with Kivimaki et al. going one step further to use Wikipedia hyperlink's to determine the skillsets.

Challenges

There were common challenges that these studies had with regards to the building of their skill models. The challenges of not having a systematic format for both JDs and CVs led to the need to develop algorithms that will interpret and add context to the skills. Gugnani and Misra took both a technical and manual approach to this challenge. They used the help of five domain experts to manually label and annotate JDs, highlighting the terms/keywords that they recognize as skill-phrases. Whereas Bastian et al. used three different technical methods to gain clarity during their skill extraction phase. This was the *Disambiguation: Clustering to provide context,* and *Deduplication: Crowdsourcing*. Similarly to Kivimaki et al., they made the use of wikipedia pages to establish definition, but also used Amazon's Mechanical Turk to provide more context to the skill. Finally, Kivimaki et al. made the use of a spread activated algorithm to tackle their specific problem. I was particularly impressed by the way these three papers tackled the challenges they came in contact with when trying to develop their skill extraction pipelines. From this literature review, the combination of manual and technical approaches (more technical approaches) will give us the most accurate result. I think the manual should be included as the final decisions of the recruitment will be based on humans, therefore I feel if we can implement the human rationale to our algorithm we will be able to get a more balanced prediction.

Skill Recommendations

A common approach following the skill extraction is providing skill recommendations. One way of making the match between JDs and CVs as accurate as possible is through the use of scoring systems. Gugnani and Misra used the system to combat a challenge they had with JDs and CVs. They noticed that JDs with more skill requirements than others will tend to have a higher matching rate with CVs, which would lead to an inaccurate and poor matching system. To combat this the introduction of an Affinity score system was introduced. This system measures the suitability of the recommendation for a CV and JD, with scoring ranging from 0 (poor match) to 1(suitable match). I particularly found this scoring system effective as it improves the accuracy of the prediction which is extremely beneficial for the recruiter.

Section 4: Successes and Weakness

Study	Strengths	Weakness
Implicit Skills Extracting using Document Embedding and Its Use in Job Recommendation by Gugnani and Misra	● The thorough use of the NLP process was essential in their approach to build their model. ● The building of a skill dictionary helped them overcome their initial challenges ● The use of Affinity score strengthens definition of skills.	● The use of manual determinations of a skill brings the use of a human element, however it is not efficient enough.
Linkedin Skills: Large-Scale Topic Extraction and Inference by Bastian et al.	● Their method to be rid of initial complications is very effective. Authors were very meticulous in their approach which improves the quality of their results. The methods reduced ambiguity, and also decluttered the keywords. ● Use of a diverse dataset.	
A Graph-Based Approach to Skill Extraction from Text by Kivimaki	● Use of a mathematical way of skill extraction	● The use of wikipedia links is flawed due to user's ability to edit pages. ● Not tested on a diverse amount of datasets, therefore their conclusion is based on limited tests. ● The use of unsupervised learning causes more complications and leads to potential inaccuracies in their model.