Representation of Job-Skill in Artificial Intelligence with Knowledge Graph Analysis

1st Shanshan Jia
Smart City College
Beijing Union University
Beijing, China
jssll00@126.com

2nd Xiaoan Liu
Smart City College
Beijing Union University
Beijing, China
liuxiaoan8008@126.com

3rd Ping Zhao
Smart City College
Beijing Union University
Beijing, China
932678068@qq.com

4th Chang Liu
College of Robotics
Beijing Union University
Beijing, China
liuchang@buu.edu.cn

5th Lianying Sun
College of Urban Rail Transit and Logistics
Beijing Union University
Beijing, China
sunlychina@163.com

6th Tao Peng *
(College of Robotics
Beijing Union University
Beijing, China
pengtao@buu.edu.cn

***Abstract*—This study analyses the relationship of different key skills of artificial intelligence (AI) used in the job market. For this, we represent it with a knowledge graph and use a Long Short-term Memory Network to study interactions between these key skills. First, a knowledge graph is build with a rule-based method about these skills in the job markets. Then, the graph is visualized to discover knowledge relationship. Jobs in AI can be classified into two categories: general algorithm jobs and specific focus jobs. Skills of different jobs in AI are very different. Python and Linux are the most necessary key skills for all jobs in AI. Revealing all these key skills in jobs in AI is useful and provide a guideline for job-seekers, companies and universities.**

***Index Terms*—Artificial Intelligence Knowledge Graph**

I. Introduction

In recent years, with the development of artificial intelligence (AI) technology, new business companies are growing in a fast fashion [1], [2]. As a result, the jobs related to AI show an explosive growth, however, the key skills used to train the people working in the AI field have not been fully studied and discussed. Therefore, the analysis of those key skills in AI through data-driven methods can guide job-seekers to clarify the skill development and goals, and provide a guild-line for enterprise hiring and school-training programs.

At present, there are few studies focus on analyzing the skills in AI job market. Liu et. al [3] used statistics and manual methods to extract key words of big data job market information, and then clustered the entities to analyze the demand for big data jobs, in which they mainly analyzed the overall demand of big data jobs from a macro perspective. Zhong et. al [4] proposed a method based on advertising content for analysis of the key components of employment, in which they used statistical analysis, variance analysis and association rule to extract information. Xia et. al [5] referred to the bibliometric method to extract the textual features of the geographic, industry and job responsibilities during the recruitment. Based on these characteristics, the relationship between academic research of big data and the needs of enterprise hiring was studied. Xu et. al [6] proposed a data-driven modeling approach that analyzed the popularity of job skills under large-scale recruitment data. Anderson et. al [7] proposed a web-based approach for measuring the skills of job seekers and using web analytics tools to analyze the demand for skills in the recruitment market. Karakatsanis [8] used latent semantic analysis to calculate the semantic similarity between job titles and job information, and then sorted and analyzed the job information from multiple aspects.

The studies mentioned above can effectively analyze the overall demand picture of the recruitment market. However, in the face of the explosive growth of AI related jobs, there are still three shortcomings in these studies. First, the data they used are very limit such that it is not enough to dig and obtain the information within the data. Second, regardless of the jobs or skills of the analysis, identifying the corresponding entity features is a top priority, and in the past studies, the most advanced methods were not used. Third, the relationship between the specific jobs and skills cannot be further analyzed.

In order to solve the problems mentioned above, we crawl a large number of the latest job requirements from recruitment websites. In addition we use Long Short Term Memory (LSTM) network to identify the entities of skills. Moreover, we build a knowledge graph about jobs and skills to discover knowledge between them. Finally, we visualize the results and do further analysis .

II. Building knowledge graph of AI jobs and skills

Knowledge graph is a semantic network that reveals the relationship between entities, which can formally describe the

This work is partially supported by grants from Beijing Municipal Education Commission Science and Technology Development Plan Project under Grant KM201211417002

Shanshan Jia and Xiaoan Liu contributed equally to this work.

*Corresponding author Email address: pengtao@buu.edu.cn (Tao Peng)

things in the real world and their interrelationships. Its basic unit includes the 'entity-relationship-entity' named resource description framework (RDF) and the entity and its related attribute-value pairs, which are interconnected by relationships to form a network of knowledge structures [9], [10]. At present, there are few studies that apply knowledge maps to study the relationship between jobs and skills, but mainly focus on smart search, deep question and answer, and social networking [11]–[13]. The natural attributes of the relationship between the constructed entities of the knowledge graph make it an inevitable tool for studying the relationship between jobs and skills.

A. Research framework

The overall research framework of this paper is shown in algorithm 1.

We used the AI related job texts on the recruitment websites and preprocessed the data to obtain high quality data. In addition, we built a job-skill knowledge graph through a deep learning-based entity recognition method, termed as a N-gram-based language model, and a rule-based relationship extraction method. Finally, we visually analyze the job-skill knowledge graph to further explore the relationship and knowledge between jobs and skills.

B. Data acquisition and preprocessing

We used python script to crawl job text data related to AI from a large set of websites, including lagou.com, zhaopin.com, and zhaopin.baidu.com. As of May 20, 2018, a total of 31258 jobs and their demand data were available. We crawled the contents including job title, job description, requirement. Due to the diversity of job text data formats and the inclusion of some useless information, we preprocessed the original job and requirement data based on rules and regular expressions, including data deweighting, symbol normalization and removing stop words, then we got a size of 21158 as useful data.

C. Job-skill knowledge graph construction

Knowledge graphs not only express the information of the Internet as a form closer to the world of human cognition, but also provide a better way to organize, manage and use massive amounts of information [14]. At first, we use LSTM to extract named entities, and then we use the semantic semantic similarity calculation method of the N-gram model to realize the job entity link. Finally, the relationship between the job and the skill is extracted by the rule method, and the information extraction results are stored and represented by the RDF triplets.

1) Skill entity recognition: Lample et. al [15] proposed an LSTM-based network for named entity recognition and achieves good results on multiple public data sets. Therefore, in this paper we uses the pre-trained LSTM network based on word granularity for skill named entity recognition. The network structure is shown in Fig 1.

The input of the model is the job requirement text, and the output is the sequence label result corresponding to the text. As shown in Fig 1, O represents a general character, E represents a skill entity character, and the middle diamond represents an LSTM unit. We find that this method can effectively extract the skill entities in the job requirement text. However, after the skill entity is extracted, there are a small number of entities that need to get further physical links, which are to map multiple expressions of one skill to the same skill. For instance, we should link CNN and Convolutional Neural Network to the skill entity CNN. Finally, we construct a skill entity mapping dictionary by manual method.

2) Linking position entity: Since there are multiple descriptions of job titles in the same job information in the recruitment website, for example, NLP, NLP Engineer and Natural Language Processing Engineer. After sorting and analyzing the job titles, we can divide all job titles into the following 10 categories: machine learning engineer, deep learning engineer, autopilot engineer, natural language processing engineer, image processing engineer, speech recognition engineer, pattern recognition engineer, computer vision engineer, robot engineer and algorithm engineer. Therefore, we use the semantic similarity calculation method based on the Bi-gram model to link the job title with same meaning to a standard job title, and then get the formalized position entity, as shown in formula (1).

$$sim(P_i, C_j) = \frac{p_i \cap c_j}{p_i \cup c_j} \tag{1}$$

where p_i represents the binary co-occurrence word set of the job name $P_i c_j$ represents the binary co-occurrence word set of the category C_j. The similarity result interval is between [0, 1], and the closer to 1 is, the more similar P_i and C_j are.

3) Position-skill relationship extraction: Through analyzing the job text data, we define three kinds of relationships: understanding, familiarity, proficient. The keywords that specifically represent these three relationships are shown in Tab I,

TABLE I
JOB AND SKILL RELATIONSHIPS AND KEYWORDS

Relationship	Key words
understand	understand ...
Familiar	familiar, have, well ...
Proficient	proficient, skillful, master ...

By using the rule-based method, the relationship between the corresponding skills and jobs is extracted. The extracted results are stored in the form of RDF triples, and some examples are shown in Tab II.

III. RESULTS

We constructed an artificial intelligence related job-skill knowledge graph (as shown in Fig 2). According to statistics, the entity with the most outreach is algorithm engineer, and the entity with the most degree of entry is Python. In order to further analyze the data of the job-skill knowledge graph, we will further analyze the knowledge graph from the perspective of jab and skill respectively.

Algorithm 1 Building knowledge graph of AI jobs and skills

Require: raw data
Ensure: rdf_format_data

```
# prepare raw data
data = get_positon_data();
new_data = text_preprocessing(data);
# load model
skill_ner_model = NER_LSTM_model (trained_model_path);
position_clf_model = position_clf(clf_model_path);
relation_extraction_model = re_model(re_model_path));
# skill entity recognition
skill_entities = position_ner_model.predict (new_data.skill_sentences);
# positon entity linking
format_positions = position_clf_model.predict (new_data.positions);
# relationship extraction
relationships =re_model.predict(new_data);
# change to rdf data format
rdf_format_data = RDF(format_positions,relationships,skill_entities);
# save and visualzation
save(rdf_format_data);
show(rdf_format_data);
```

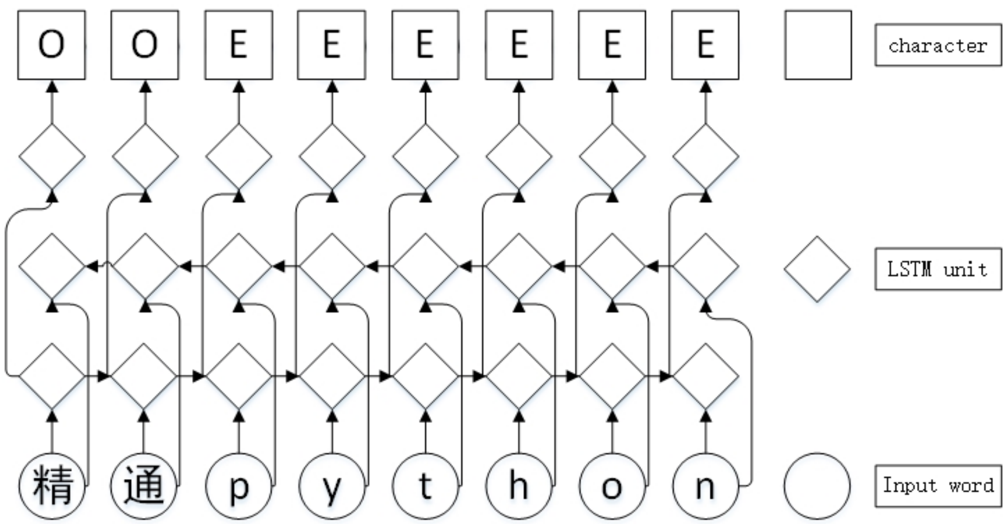

Fig. 1. Network structure of named entity recognition.

TABLE II
JAB AND SKILL RDF EXAMPLES

position-skill RDF
machine learning, proficient, scipy
machine learning, proficient, numpy
machine learning, familiar, keras
Natural language processing, familiar, NLP
Image, familiar, spark

A. Results from the perspective of the job

We divide all jobs into 10 categories, and the recruitment enthusiasm and proportion of each job are shown in Fig 3,

As we can see from Fig 3, the jobs related to artificial intelligence are mainly divided into two categories. One is the job oriented to the basic algorithm which has a dominant position, including machine learning engineer, algorithm engineer, deep learning engineer. The other category is for specific research areas, including natural language processing engineers, image processing engineers, speech recognition engineers, etc. In the specific research field, it can be found that the recruitment needs are mainly concentrated in the three main media of image, text and speech. Further, we visualize the position-skills map for each position. Due to the small number of pattern recognition and robotic engineer recruitment information, we

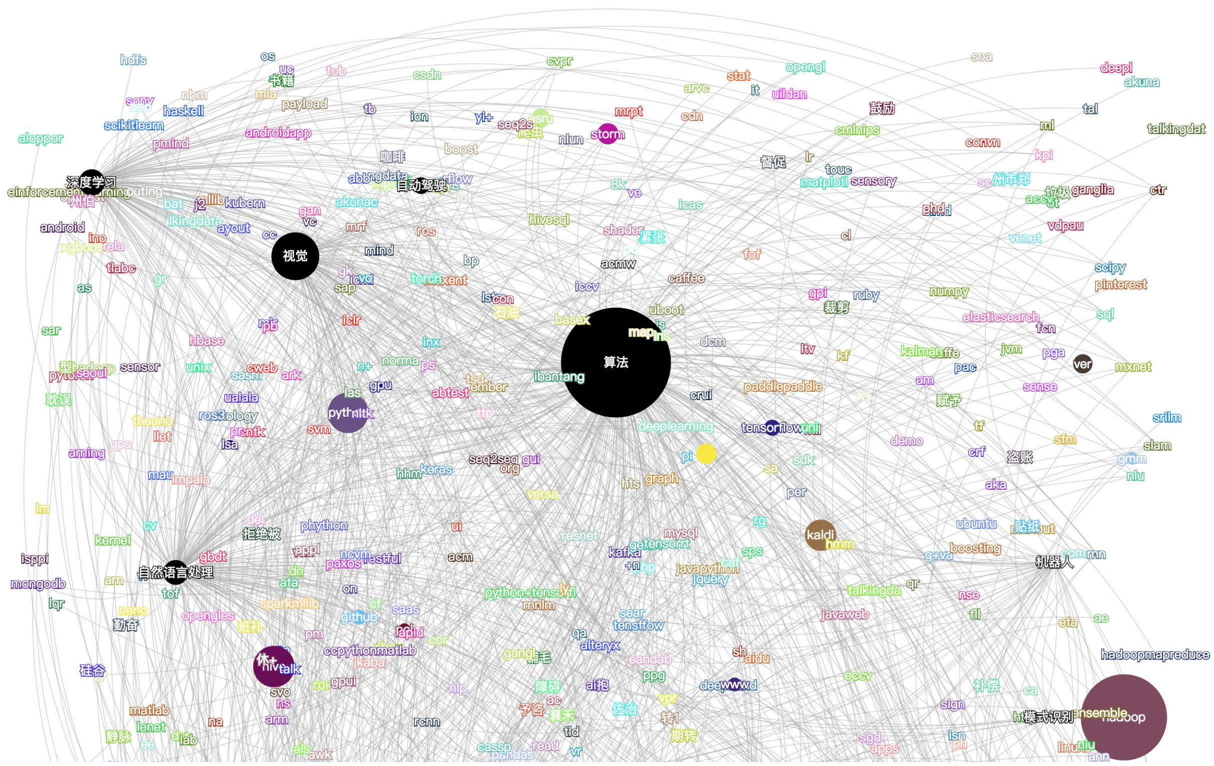

Fig. 2. Job and skill knowledge graph.

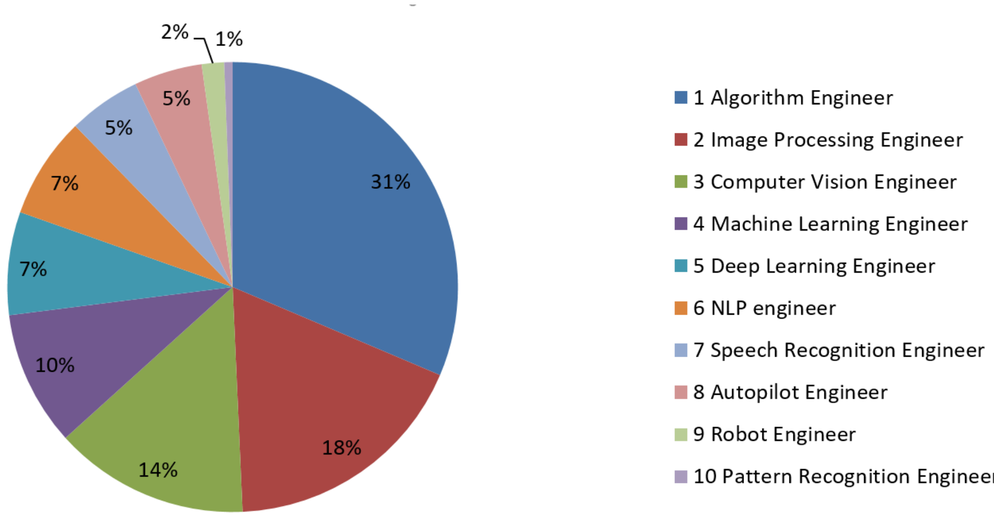

Fig. 3. Job and skill RDF examples

focus on the other eight categories. Through the analysis of the map, the requirements for skills in different positions is vary dramatically.

- The requirements for the skills of the algorithm engineers and machine learning engineers mainly include programming languages such as Python, Java, and C++. In addition, there are clear requirements for big data processing frameworks such as Hadoop and Spark which means that these two types of positions not only focus on the engineer's algorithmic programming capabilities, but also on the ability of engineers to handle big data. encouraged.
- The requirements for skills in deep learning engineer positions include Caffe, Tensorflow, Python, CNN, RNN, CUDA, GPU, etc. It can be seen that for deep learning engineer position, the company's skill requirements are

TABLE III
NUMBER OF SKILL-RELATED POSITION CATEGORIES

Skills	Number of associated positions	Skills	Number of associated positions
Python	10	Matlab	8
Linux	10	Torch	8
C++	9	RNN	8
GO	8	CNN	8
Java	8	Shell	8
Tensorflow	8	Android	8
Caffe	8	DNN	7

clearly focused on various frameworks related to deep learning and classical neural network models.

- Natural language processing engineers' skills requirements in programming languages, deep learning frameworks, neural network models, etc. It is also necessary to focus on models such as Latent Dirichlet Allocation (LDA), Conditional Random Field (CRF), and Support Vector Machine (SVM).
- Different from the previous positions, speech recognition engineers also need to master the mainstream tools of speech recognition (such as: Kaldi).
- Computer vision engineers' skills requirements mainly include Opencv, C++, Python, Caffe, etc. To be further analysis, we find that the talents the company needs are more focused on mastering the computer vision library, and secondly on the mainstream programming language and deep learning framework.
- The skills that robot engineers need to master are mainly ROS, Linux, Python, SLAM, PID, Kalman, which means that robotics engineers need to focus on the proficiency of their operating systems such as Linux, as well as the skills of the real-time map construction, controller and Kalman filtering methods.

B. Results from skill analysis

We analyze the relationship between jobs and skills from skill perspective. According to the job correlation of each skill, the statistics are sorted. The results are shown in Tab III. As can be seen from Tab III, the current skill requirements for artificial intelligence related jobs are mainly concentrated in three aspects: programming languages (such as Python), deep learning frameworks (such as Tensorflow), and classical neural networks (such as CNN).

As can be seen from the associated job categories, Python and Linux are required for all AI-related positions. Python is a scripting language widely used in artificial intelligence related positions because of its simple and flexible code style and rich and powerful third-party libraries.

- Linux is a free, stable and widely used server operating system. Because most of the existing artificial intelligence related work is based on large data volume and large computing power, which makes the engineers need to train and test their models through the server, Linux has become a must-have skill.
- In addition to Python, C++, Java and other mainstream programming languages also play an important role in the field of artificial intelligence.
- Moreover, GO is also widely used due to its powerful standard library and fast compilation speed.
- In the deep learning framework, the requirements for artificial intelligence related positions are mainly concentrated in Tensorflow, Caffe, Torch, Theano, Pytorch, MXNet. Further analysis of job text data found that there is no clear-cut learning framework that needs to be understood, and just ask for one of them. This shows that the existing deep learning frameworks are still in a state of balanced development, which has no significant gap in the level of recognition of the major frameworks in the industry.

IV. CONCLUSIONS

We used LSTM model for skill-named entity recognition, Bi-gram-related algorithms for job entity linking, RDF triples for storing and representing information extraction results, and processing artificial intelligence related job text data to build artificial intelligence job-skills knowledge graph.

By analyzing the knowledge graph from the perspective of job and skill respectively, we found that the jobs are mainly divided into basic algorithm jobs and specific research field jobs. Due to the different skill requirements in different positions, job seekers should learn specific skill according to the specific position, and Python and Linux are essential skills for AI positions.

Our analysis results have certain reference values for job seekers, companies and universities in the field of artificial intelligence. In addition, the research ideas of constructing the job-skill knowledge graph in this paper can be further applied to other industries.

In the further, we will build a more comprehensive job-skills map based on this paper, taking other types of entities and relationships such as salary, company size and geographic location into account. In addition, the study can be further extended, such as combining with the resume of job seekers, to build a smart resume screening and resume recommendation system based on knowledge graph.

ACKNOWLEDGMENT

We would like to thank Zhaofei Yu for helpful discussions.

REFERENCES

[1] Zhu C, Zhu H, Xiong H, et al. Recruitment market trend analysis with sequential latent variable models[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 383-392.

[2] Gardiner A, Aasheim C, Rutner P, et al. Skill Requirements in Big Data: A Content Analysis of Job Advertisements[J]. Journal of Computer Information Systems, 2017: 1-11.

[3] Liu Ruilun, Ye Wenhao, Gao Ruiqing, et al. Research on Text Clustering Based on Requirements of Big Data Jobs[J]. Data Analysis and Knowledge Discovery, 2017, 1(12):32-40.

[4] Zhong Lihong, Deng Zhihong.Research on Requirement Method Based on Requirements Advertisement Content[J]. Journal of Shenzhen Institute of Information Technology, 2017(4):85-92.

[5] Xia Huosong, Pan Xiaoting. Research on Relationship Between Big Datas Academic Research and Its Talent Demand Based on Python[J]. Journal of Information Resources Management, 2017, 7(1): 4-12.

[6] Xu T, Zhu H, Zhu C, et al. Measuring the Popularity of Job Skills in Recruitment Market: A Multi-Criteria Approach[J]. 2017.

[7] Anderson K A. Skill networks and measures of complex human capital:[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(48):12720.

[8]] Karakatsanis I, AlKhader W, MacCrory F, et al. Data mining approach to monitoring the requirements of the job market: A case study[J]. Information Systems, 2017, 65: 1-6.

[9] Xu Zenglin, Sheng stroke pan, He Lirong, et al. Review on Knowledge Graph Techniques[J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4):589-606.

[10] Liu Jiao, Li Yang, Duan Hong, et al. Knowledge graph construction techniques[J]. Computer research and development, 2016, 53(3):582-600.

[11] AMIT S. Introducing the knowledge graph[R]. America:Official Blog of Google, 2012.

[12] Lukovnikov D, Fischer A, Lehmann J. Neural Network-based Question Answering over Knowledge Graphs on Word and Character Level[C]// International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017:1211-1220.

[13] Cao Qian, Zhao Yiming. The Realization Process and Related Applications of Knowledge Graph[J]. Information Studies: Theory and Application(ITA), 2015, 12(38): 127- 132.

[14] Li Juanzi, Hou Lei. Review on Knowledge Graph Research[J]. Journal of Shanxi University (NATURAL SCIENCE EDITION), 2017, 40(3):454-459.

[15] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.