2021 2nd International Conference on Big Data and Informatization Education (ICBDIE) | 978-1-6654-3870-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICBDIE52740.2021.00023

Skill Requirements Analysis for Data Analysts Based on Named Entities Recognition

Lina Cao
School of Economics and Management, Beijing International Science and Technology Cooperation Base of Intelligent Decision and Big Data Application
Beijing Information Science and Technology University
Beijing, China
caolina@amss.ac.cn

Jian Zhang*
School of Economics and Management, Laboratory of Big Data Decision Making for Green Development
Beijing Information Science and Technology University
Beijing, China
* Corresponding author: zhangjian@bistu.edu.cn

***Abstract*—Currently, with the surge in demand for data analysts, it is beneficial to grasp the social requirement for talents by deeply mining the job requirements of data science. This study used crawlers to acquire the job advertisements from 51 job.com and selected data science related positions. Then, the specialty and skill entities in job requirement were annotated manually, and two kinds of Named Entities Recognition models, the Bert-BiLSTM-CRF model and BiLSTM-CRF were applied to extract skill entities and specialty entities from the text data of job requirement. It was found that, the model based on Bert pre-training vector was better. At last, the characteristics of the demand from the perspective of skills and profession were analyzed.**

Keywords-component; data science; job advertisement; Named Entities Recognition

I. Introduction

The advent of the data era has brought great changes to all walks of life, and enterprises have an increasingly urgent need for data analysts. On February 16, 2016, Data Science and Big Data Analysis Major was added for the first time in the 2015 Annual Record and Approval Results of Undergraduate Majors in Colleges and Universities released by the Ministry of Education. Chinese colleges and universities began to set this specialty for undergraduates. However, this major has been opened for a short time, and the teaching system of this subject is still in the process of updating, upgrading and improving. As an emerging major, how to better match the needs of enterprises and train big data talents to meet the needs of society has become an important issue related to the development of this major.

Recruitment websites are the main data source reflecting the social demand. By mining talent demand characteristics and designing the talent training scheme can improve the suitability of the talent supply and the demand of enterprise, adapt to the new requirements of personnel training under the background of new economy, new industries and new technology.

II. Review

The core technology of extracting demand information from data science-related job postings is information extraction technological which can be divided into the following kinds according to the development process. Only then can the extracted information be further analyzed and summarized.

There were many researches on the skill extraction. The method based on dictionary was the earlier method. For example, Hämäläinen et al. [1] and Aken & Litecky [2] constructed a skill dictionary or list and matched the dictionary with the recruitment text information to recognize skill keywords. The method based on statistics is to sort keywords by feature quantitative indicators. Grüger & Schneider [3] used TF-IDF values to filter skill keywords, and this methods do not require syntactic and semantic information, and do not rely on labeling data, but the accuracy rate is relatively low. The method based on probabilistic topic models is to generate skills-related topics. For example, Latent Dirichlet Allocation (LDA) model was used the to obtain potential topic-skill distributions [4]. The hybrid method based on rules and statistics can improve the performance of keyword extraction. For example, the CareerBuilder [5] developed a SKILL system via a word vector model to realize skill entity recognition and entity specification. However, these methods still need to be improved in terms of accuracy or efficiency.

Named Entity Recognition (NER) is one of the important tasks in information extraction and information retrieval. It can be used to identify and classify named entities in text, such as names of people, places and institutions [6,7]. The combination of deep learning and NER is also being implemented by more and more researchers. Deep learning algorithms, from the classical RNN, CNN, Long Short-Term Memory (LSTM), Bi-directional Long Short-Term Memory (BiLSTM), etc., are constantly injecting new vitality into NER. Through deep data processing, corresponding features are extracted layer by layer to improve the efficiency of knowledge aggregation [8]. So far, there has been no attempt to apply BERT to Named Entity Recognition in job ads. This paper is an exploratory study trying to use BERT-BiLSTM-CRF to extract skill information.

III. Model description

The sequence labeling of BERT-BILSTM-CRF model can be roughly divided into three stages: word vector representation based on Bert, context feature learning based on BiLSTM and maximum label sequence output based on CRF.

A. Bert

BERT is a preliminary training language representation model proposed by Devlin [9]. It is not like Peters and Radford [10, 11] as the traditional language models which train from left to right or from right to left. Instead, it innovative uses the

978-1-6654-3870-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICBDIE52740.2021.00023

masked language model (MLM) and adds a loss that predicts the next sentence. The MLM task is to train deep bidirectional features so that word representations can better integrate the context of the context. The next sentence prediction task is designed to enable the model to understand the relationships between sentences. In this way, the word vector obtained by BERT not only implies the features between contextual words, but also captures some sentence-level features.

B. *BiLSTM-CRF*

LSTM [12] is a variant form of cyclic neural network RNN. BiLSTM networks [13] are composed of a forward and backward LSTM, and can be seen as a two-layer neural network with forward and backward input sequences, respectively. Conditional Random Field (CRF) is the last part of the model, which is responsible for capturing the dependency between contextual tags and constraining the contextual tags.

IV. Experiment design

A. *Web crawler got data*

51job.com is one of the largest online recruitment platforms in China. This paper used a web crawler developed by Python to capture all the job advertisement of 51job.com in June, 2020, with a total number of 3,655,953. Each piece of data included nearly 20 dimensions such as recruitment position name, educational requirement, years of experience requirement, job responsibility requirement, employment requirement, recruitment enterprise type, and enterprise size. Then search the positions with the keywords “data science” and “data mining”. After noise elimination of data, there are 5,533 pieces of ads selected.

B. *Named Entities Recognition*

Entities of skill requirements and specialty requirements were taken as named entities. The BILSTM-CRF model and Bert-BILSTM-CRF model were constructed respectively and the effect of model train were compared.

At first, we used the BIOES scheme for text annotation, where B represents the beginning of an entity, I represents the inside, E represents the end, S represents a single entity, and O represents the out. More than 5,000 pieces of ads were randomly selected from all data for manual annotation and more than 60,000 entities was labeled in the experiment.

At second, the 5000 sample data were divided into training data set, validation data set and test data set according to the ratio of 8:1:1. The model performance was tested by 10-fold cross-validation, and the effect of the model was assessed by precision rate, recall rate and F-score. An entity was considered to be correctly labeled only if its type and start-stop boundary were both correctly. The experiment results of the two models are shown in Table 1.

TABLE I. Comparison of entity recognition results of the models

Evaluation index Named entities	BiLSTM-CRF model			Bert-BiLSTM-CRF model		
	Precision Rate（%）	Recall Rate（%）	F-score（%）	Precision Rate（%）	Recall Rate（%）	F-score（%）
Specialty entity	93.18	83.69	85.01	95.56	85.10	86.65
Skill entity	84.32	86.22	85.89	86.40	87.54	86.76

As the Table 1 shown, compared with the BILSTM-CRF model, the recognition effect of the model based on Bert pre-training vector was better. This model with higher accuracy was selected to realize the automatic extraction of named entities in the job ads related to data analysis.

V. Experiment result

After the prediction of the model trained by Bert-BiLSTM-CRF, the specialty and skill entities were recognized.

A. *Specialty entities analysis*

1) Popular specialties requirement analysis. According to the experimental results and statistics, there were 3614 pieces of online job postings were extracted specific required majors. Counting the frequency of required specialty entities, and then calculating and sorting the proportion of each major in all pieces of occupational data, the popular majors required by data analysis and mining would be obtained as shown in Figure 1. It indicates that the most desirable majors for this vocation are Statistics, Mathematics and Computer Science, which are mentioned respectively in 60%, 50% and 50% of the 5,533 data samples. There are also other majors such as Economics, Finance, Financial Accounting and Marketing required when considered the application scenarios or industries of data analysis and process. Unexpectedly, the major of Medicine and Environment rank eighth and tenth, respectively, due to the emerging demands of medical and environmental data analysis.

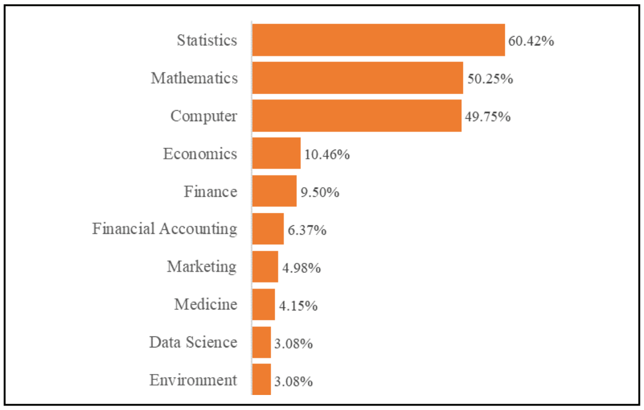

Figure 1. The order of ratio of professions requirement

Data Science, which should be the most appropriate major for these positions, however ranked ninth among all specialties. Actually, 2020 was the year when the first students graduate. Obviously, according to the data result, the social cognition degree of this major is not high.

2) Specialties co-occurrence analysis. To reveal the interdisciplinary knowledge required by the positions and the span of specialized knowledge to be mastered, the co-occurrence frequency of the required specialty entities in each recruitment text was counted. For example, if a job advertisement text mentioned the need for major of Statistics and Computer, the co-occurrence frequency between Statistics and Computer major would be recorded as one. After counting all co-occurrence frequencies, majors with co-occurrence frequency greater than 10 in pairs were taken as nodes, and co-occurrence frequency was taken as the weight of the connecting edge of the two majors, then the co-occurrence graph of majors would be draw out as shown in Figure 2 below.

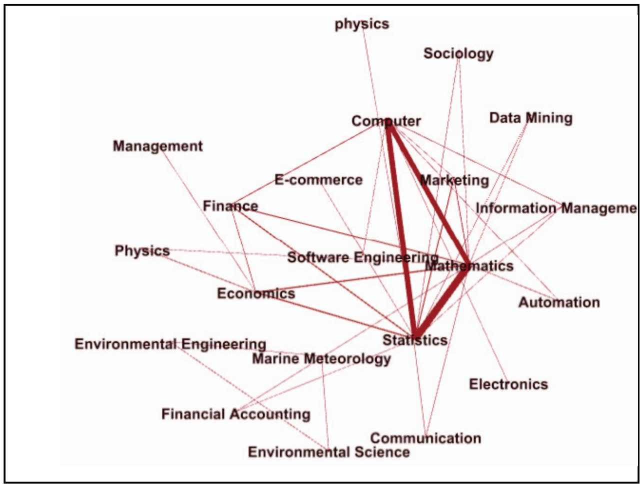

Figure 2. Map of the co-occurrence of professions

As can be seen from the Figure 2, Statistics, Mathematics and Computer have the highest co-occurrence frequency, indicating that most enterprises basically choose these three majors when recruiting the data analysis and processing engineering technicians, which is consistent with the conclusion drawn from the analysis in 1) above. Obviously, for those employees major in three specialties, mastery of the other two specialized knowledge and skills will be helpful when they engaged in the data analyst. Some enterprises may require one or two of three majors combined with other application-oriented majors, such as Marketing, Finance, Economics and so on considering the application background of data analysis and processing. For those students majoring in the application-oriented specialties, only industry knowledge will not be competitive, and they must also have some solid math knowledge or master computer tools to be better qualified the job.

B. Skill entities analysis

The statistical results of the extraction of skill words were shown as the cloud chart in Figure 3 and Figure 4. Figure 3 shows the top 30 words of core skills (general skills) required by data analyst. The abilities mentioned most mainly focused on communication ability, sense of responsibility, logical thinking, teamwork spirit, responsibility and other soft skills. Figure 4 shows the top 40 professional skills required by the occupation, including data analysis ability, data sensitivity, and data analysis tools such as Excel, SQL, Python, and SPSS. Therefore, from the perspective of labor supply side, the general skills and professional skills of the workers are equally important.

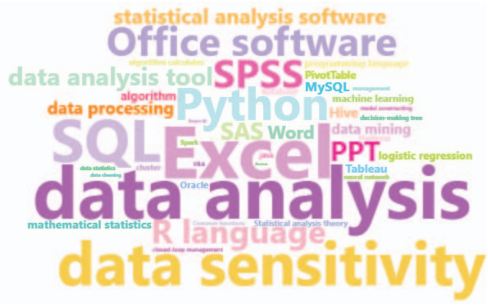

Figure 3. Cloud map of core skills words

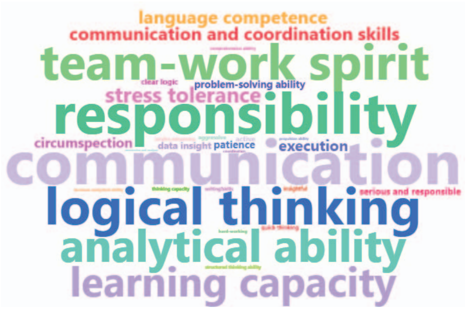

Figure 4. Cloud map of professional skills words

In the job ads, positions can be summarized to three kinds which are data analysis, big data mining and data administration. To analyze the difference of those positions, the professional skill words were measured and ordered according to the position categories as it shown in Table 2.

TABLE II. THE ORDER OF PROFESSIONAL SKILL WORDS IN EACH POSITION

data analysis	Big data mining	Data administration
analysis ability	analysis ability	analysis ability
excel	Python	excel
SQL	SQL	office
office	R	PPT
PPT	data mining	data sensitivity
data sensitivity	machine learning	SQL
Python	hive	Python
SPSS	excel	software
R	logistic regression	R
data processing	SPSS	data processing
software	SPARK	data analysis tool
word	Hadoop	statistical analysis
data analysis tool	Java	MySQL
SAS	data sensitivity	word
data mining	programming language	machine learning

Table 2 shows that those positions all require the data analysis ability, data sensitivity and office software but give different weight to the most frequently used method of statistical analysis software, programming language and database management software. Also it indicates the varying degree of difficulty in engaging in different jobs. The big data mining position has the highest requirement whether from the number of skills or the terms of professional skills. Jobs of data analysis need roughly the same number of skills, but the latter emphasize database capabilities. Jobs of data administration require less skills but comprehensive abilities.

C. Other dimension analysis

1) Degree requirements analysis. Statistics on the degree requirements of the data analyst showed that, compared to the average degree demand distribution which was the statistical result of all the data, this occupation had a higher degree requirement. The specific distribution of them was shown in Figure 5: in the data set, except for 5% of the job posting data without the educational requirement, 64% of the data requires bachelor's degree or above, which is about 38 percentage points higher than the average; the proportion of college degree requirement is 28%, 17 percentage points lower than the average; the proportion requiring technical secondary degree and high school or below is 2%, 14 percentage points lower than the average. Relatively speaking, this job has a higher threshold of employment.

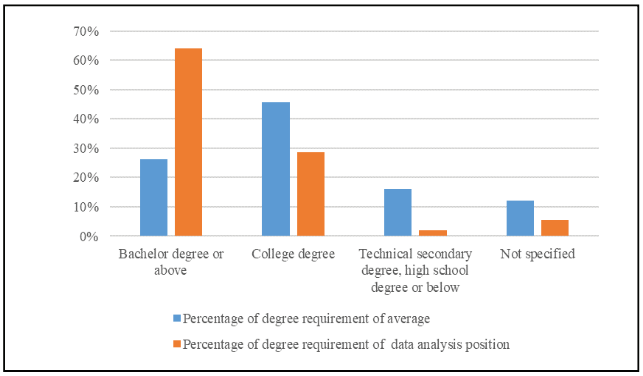

Figure 5. Comparison of distributions of the degree requirement

2) Experience requirements analysis. According to the statistics of work experience requirements, it was found that compared with the average distribution of work experience requirements which was the statistical result of all the data, the work experience requirements of this occupation were relatively high. The specific distribution was shown in Figure 6: in the data set, in addition to 8% of the sample data without work experience requirements, only 17% of the job postings data require no prior experience, which is 7 percentage points lower than the average; 42% of the sample data required 1-2 years of experience, 4 percentage points higher than the average; 25% of the sample data require 3-4 years of experience, which is 8 percentage points higher than the average; 8% of the sample data require 5 or more years of experience, 2 percentage points below the average. To sum up, three quarters of the job postings need the applicants with 1-4 years of work experience. Thus, those students who are interested in the vocation should also take into account the possible internships of the similar jobs, especially for students majoring in Statistics, Mathematics or Computer Science.

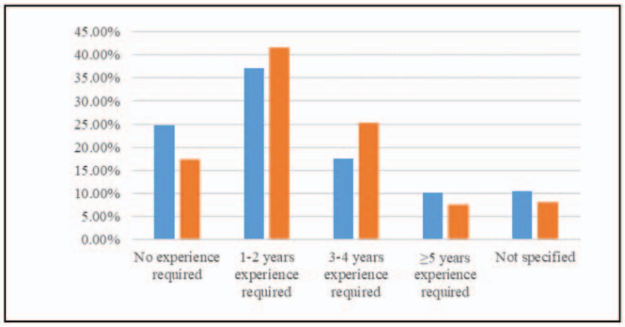

Figure 6. Comparison of distributions of the work experience requirement

VI. CONCLUSION

In this paper, in order to mining the requirements of data analyst, we analyzed the job advertisements related to data analysis positions and find the characteristic of the skill, specialty and other requirement of this job. To sum up, data analyst usually are major in Statistics, Computer or Mathematics, and core skills they commonly have are communication ability, sense of responsibility, logical thinking, teamwork spirit, responsibility, and so on, meanwhile, they should master the professional skills including data analysis ability, data sensitivity and data analysis tools such as Excel, SQL, Python, SPSS and so on. 64% of them have a bachelor's degree or above, and three quarters of them have 1-4 years of work experience.

ACKNOWLEDGMENT

Cao and Zhang are grateful to all the text annotators who promote the experiments go through.

REFERENCES

[1] H. Hämäläinen, J. Ikonen, and J. Porras, "A tool for visualizing skill requirements in ICT job advertisements", in Proc. 7th e-Learning and the Knowledge Society, Bucharest, Romania, 2011, pp. 254-259.

[2] A. Aken, C. Litecky, A. Ahmad, and J. Nelson, "Mining for computing jobs", IEEE software, vol. 27, no. 1, pp. 78-85, 2010.

[3] J. Grüger, G. J. Schneider, "Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements", in Proc. 15th WEBIST, Vienna, Austria, 2019, pp. 226-233.

[4] F. Gurcan, N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling", IEEE Access, vol. 7, pp. 82541-82552, 2019.

[5] Z. Meng, J. Faizan, J. Ferosh, and M. Matt, "SKILL: A System for Skill Identification and Normalization", in Proc. 27th IAAI, Texas, USA, March, 2015, pp. 4012-4017.

[6] N. Kabra, P. Bhattacharya, S. Tanwar, et al., "Mudrachain: Blockchain-based framework for automated cheque clearance in financial institutions", Future Generation Computer Systems, vol. 102, pp. 574-587, 2020.

[7] T. Inui, Y. Nakano, "An Analysis of Japanese Named Entity Recognizer Specialized for Person and Organization Entities", in Proceedings of 2018 International Conference on Asian Language Processing (IALP), IEEE, pp.184-188, 2018.

[8] A. Goyal, V. Gupta, M. Kumar, “Recent named entity recognition and classification techniques: a systematic review”, Computer Science Review, vol. 29, pp. 21-43, 2018.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.

[10] M. E. Peters, M. Neumann, M. Iyyer, et al, “Deep contextualized word representations”, in Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA: ACL, pp. 2227-2237, 2018.

[11] A. Radford, K. Narasimhan, T. Salimans, et al., “Improving language understanding with unsupervised learning [EB/OL], 2018, https://openai.com/blog/language-unsupervised.

[12] K. Greff, R. K. Srivastava, J. Koutník, et al., “LSTM: A search space odyssey,” IEEE transactions on neural networks and learning systems, vol. 28, no. 10, pp. 2222-2232, 2016.

[13] A. Graves, J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”, *Neural Network*, vol. 18, no. 5–6, pp. 602-610, 2005.