Automatic Extraction of Skill Information from Online Recruitment Texts: Postprint

Authors: Yú Yǎn, Chen Lei, Jiang Jinde, Zhao Naixuan

Date: 2023-07-26T00:00:00+00:00

Abstract

[目的/意义]To address the issue that manual extraction of skill information from online recruitment texts fails to meet the requirements of large-scale data analysis, this paper proposes an automatic method for extracting skill information from massive online recruitment texts. [方法/过程]Based on the characteristics of online recruitment texts, dependency parsing is employed to select candidate skills. Subsequently, a domain relevance metric is proposed to evaluate candidate skills, which is integrated into traditional term extraction methods, forming an automatic skill information extraction method for online recruitment texts. [结果/结论]Experimental results demonstrate that the proposed method can automatically, rapidly, and accurately extract skill information from online recruitment texts.

Full Text

Preamble

Yu Yan[1],[2], Chen Lei[1], Jiang Jinde[3], Zhao Naixun[1]

[1]Nanjing Tech University, Nanjing 210009
[2]Department of Computer Engineering, Chengxian College, Southeast University, Nanjing 211816
[3]School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract

[Purpose/Significance] Aiming at the problem that current manual skill information extraction from online recruitment texts cannot meet the requirements of large-scale data analysis, this paper proposes an automatic skill information extraction method for large volumes of online recruitment texts. **[Method/Process]** According to the characteristics of online recruitment

texts, dependency syntax analysis is used to select candidate skills, and then domain relevance indicators are proposed to measure candidate skills, which are integrated into traditional terminology extraction methods to form an automatic skill information extraction method for online recruitment texts. [**Results/Conclusion**] Experiments show that the proposed method can automatically, quickly, and accurately extract skill information from massive online recruitment texts.

2.2 Terminology Extraction

Terminology extraction refers to the process of automatically discovering terms from text. Currently, terminology extraction methods can be divided into unsupervised and supervised approaches. Unsupervised methods typically combine linguistics and statistics, offering advantages such as minimal manual intervention, strong applicability, and consistency. Supervised methods employ machine learning techniques to construct models for term extraction by learning features from training texts, which can compensate for the inability of unsupervised methods to recognize low-frequency terms and achieve higher precision and recall. However, they require large-scale manually annotated corpora for training, and the methods are not yet mature, requiring more attempts and validation. Since there is currently no large-scale annotated corpus for online recruitment skill information extraction tasks, this paper focuses on unsupervised methods.

Unsupervised methods typically first select candidate terms from a corpus, then use statistical information to calculate the likelihood of candidate terms becoming actual terms. Generally, termhood and unithood are used to measure this likelihood. Termhood measures a candidate term's ability to express domain knowledge, while unithood measures the structural stability of a candidate term. Specifically, the C-value method is a simple and efficient terminology extraction approach based on termhood, with numerous applications both domestically and internationally. However, since the C-value method primarily relies on the frequency of word strings in the corpus, it cannot effectively filter out non-term strings that appear frequently.

To address this issue, a typical approach is to introduce two statistical measures—mutual information and adjacency entropy—to reconstruct the C-value objective function. Mutual information calculates the dependency degree among words in a candidate term; the larger the mutual information value, the greater the dependency among words, and the more likely it is to be a term. Adjacency entropy measures the uncertainty of words adjacent to the candidate term; the greater the uncertainty, the more information its adjacent words contain, and the more likely it is to be a term. However, in online recruitment texts, some non-skill strings frequently co-occur and have high mutual information values, such as "相关专业" (related major) and "工作经验" (work experience). Therefore, mutual information cannot effectively measure candidate terms. Similarly, some high-frequency non-skill strings in online recruitment texts have high adjacency entropy, such as "熟练使用" (proficient

in using) and “具有良好” (have good), which also cannot effectively measure candidate terms.

Overall, terminology extraction research has achieved certain results, but directly applying these methods to skill information extraction from online recruitment texts would result in low precision and recall. Existing studies typically use continuous noun and verb strings to select candidate skills. However, this approach includes a large number of noisy verb-containing non-skill strings, such as “熟练使用” and “熟悉 HTTP 协议” (familiar with HTTP protocol), resulting in low final skill extraction precision. If candidate skill strings containing verbs are excluded, some candidate skills might be missed, such as “优化” (optimize) in “SQL 优化” (SQL optimization), leading to low recall.

3 Automatic Skill Information Extraction Method for Online Recruitment Texts

Based on the characteristics of online recruitment texts, this paper first uses dependency syntax analysis to select candidate skills, then proposes the concept of skill domain relevance. By introducing non-target domain online recruitment text sets based on the target domain set, we measure the domain relevance of skill information to improve the C-value method. The process is shown in Figure 2 [Figure 2: see original paper], which mainly includes preprocessing (Section 3.1), candidate skill selection based on dependency syntax analysis (Section 3.2), C-value calculation (Section 3.3), domain relevance measurement (Section 3.4), and C-value calculation incorporating domain relevance (Section 3.5).

3.1 Preprocessing

Since recruitment texts are unstructured web pages containing not only required information such as skills but also large amounts of noise such as advertisements, images, animations, irrelevant hyperlinks, scripting languages, and various tags, we first use web text analysis tools like BeautifulSoup to locate and parse web content to obtain skill requirement texts. Then, we perform deduplication, case conversion for English text, and removal of special characters on the obtained texts. Figure 3 [Figure 3: see original paper] shows an example of online recruitment text preprocessing.

3.2 Candidate Skill Selection Based on Dependency Syntax Analysis

Analysis of online recruitment texts reveals that skill-containing text typically appears in verb-object structures, such as “熟悉关系型数据库” (familiar with relational databases). Therefore, this paper proposes using dependency syntax analysis to eliminate noisy verbs like “熟悉” (familiar). Dependency syntax analysis reveals semantic modification relationships between words within a sentence unit through dependency relations, represented by directed arcs from the head word to its dependent. According to dependency grammar axioms, dependency

syntax analysis hierarchically structures the linear structure of a sentence into a dependency tree.

Figure 4 [Figure 4: see original paper] shows dependency trees T1, T2, T3, and T4 obtained by analyzing the sentences "熟悉关系型数据库" (familiar with relational databases), "并有一定的 SQL 优化经验" (and have some SQL optimization experience), "熟悉 HTTP 协议" (familiar with HTTP protocol), and "具有良好的团队合作意识" (have good team collaboration awareness) using the dependency parser released by the Harbin Institute of Technology Language Technology Platform. In Figure 4, Root points to the core verb of each sentence, and letters under nodes indicate part-of-speech tags: v for verb, n for noun, c for conjunction, b for distinguishing word, u for auxiliary, ws for foreign word, and a for adjective.

Accordingly, this paper removes the core verb pointed to by Root, retains all remaining verbs, nouns, foreign words, and other content words, and uses an n-gram strategy to select word strings with frequency greater than 1 and length between 1-4 as candidate skills. Table 1 compares candidate skills selected using dependency syntax analysis with traditional methods, where core verbs pointed to by Root in the dependency analysis are shown in bold. As shown in Table 1, the dependency syntax analysis-based method yields candidate skills containing fewer noisy strings and effectively filters out words like "熟悉" (familiar), "具有" (have), and "有" (have).

3.3 C-value Calculation

The C-value method calculates termhood for each candidate skill based on statistical information including the candidate skill's frequency, length, and the frequency and count of longer candidate terms containing it. The C-value calculation method is shown in formula (1):

$$C\text{-value}(x) = \begin{cases} \log_2 |x| \cdot tf^{(T)}(x) & \text{if } x \text{ is not nested} \\ \log_2 |x| \cdot \left(tf^{(T)}(x) - \frac{1}{|C_x|} \sum_{y \in C_x} tf^{(T)}(y) \right) & \text{otherwise} \end{cases}$$

where x represents a candidate skill, $|x|$ denotes the length of x, $tf^{(T)}(x)$ indicates the frequency of x in the target online recruitment text set T, C_x represents the set of candidate skills containing x in the target set, and $|C_x|$ denotes the number of elements in set C_x. Formula (1) shows that C-value is related to the candidate skill's frequency in the target corpus—the higher the frequency, the greater its termhood. Additionally, it considers the length of candidate skills, as the frequency of longer strings is considered more meaningful than that of shorter strings.

3.4 Domain Relevance Measurement

To measure the domain relevance of candidate skills, this paper first measures the domain relevance of each word in the candidate skill, then calculates the domain relevance of the entire candidate skill string based on the relevance of its constituent words.

3.4.1 Word Domain Relevance Measurement Skill information consists of several words. This paper proposes word domain relevance (DR) to describe the degree of association between a word and a specific domain. Specifically, given a target domain online recruitment text set T and a non-target domain online recruitment text set NT, we compare the frequency of word w in T and NT, defined as formula (2):

$$DR_w^{(T)} = \frac{p_w^{(T)}}{p_w^{(NT)}} = \frac{tf_w^{(T)}/|T|}{tf_w^{(NT)}/|NT|}$$

where $p_w^{(T)}$ represents the probability of word w appearing in the target domain set T, $tf_w^{(T)}$ denotes the frequency of word w in T, and $|T|$ represents the total number of words in the target domain set. Similarly, $p_w^{(NT)}$ represents the probability of word w appearing in the non-target domain set NT, $tf_w^{(NT)}$ denotes the frequency of word w in NT, and $|NT|$ represents the total number of words in the non-target domain set. Formula (2) shows that a larger DR value indicates stronger relevance to the target domain, while a smaller DR value indicates weaker relevance.

3.4.2 Candidate Skill Domain Relevance Measurement A candidate skill contains several words $x = \{w_1, w_2, ..., w_m\}$. Based on word domain relevance, we measure the degree of relevance between the candidate skill string and a specific domain. The calculation method is shown in formula (3):

$$DR_x^{(T)} = \prod_{i=1}^{m} DR_{w_i}^{(T)}$$

where $DR_x^{(T)}$ represents the domain relevance degree of candidate skill x in the target domain T. By definition, only when every word in the candidate skill has high domain relevance will the candidate skill itself have high domain relevance.

3.5 C-value Calculation Incorporating Domain Relevance

When a candidate skill has a larger C-value and greater domain relevance, it is more likely to be a skill. Therefore, this paper proposes C-value calculation incorporating domain relevance to measure the likelihood of a candidate term being a skill. The calculation method is shown in formula (4):

$$DRC\text{-value}(x) = DR_x^{(T)} \times C\text{-value}(x)$$

Finally, the DRC-values are sorted in descending order, and the top candidate skills are selected as extracted skills.

4 Experiments

4.1 Dataset

To verify the feasibility and effectiveness of the proposed method, we crawled recruitment text data from the mainstream Chinese recruitment website "51job.com" (www.51job.com) to extract skill information from computer domain recruitment texts. 51job is a leading online recruitment service provider in China. We selected data from the "Computer/Internet/Communication/Electronics" function as the target domain online recruitment text set, with data crawled from March 19, 2018 to March 26, 2018. We also crawled data from other non-computer-related functions on 51job during the same period. After removing recruitment texts below bachelor's degree, duplicate content, all-English texts, and those without job requirements, the basic dataset information is shown in Table 2 .

4.2 Experimental Steps and Evaluation Criteria

The experiment first preprocesses both target and non-target domain online recruitment text sets. Dependency syntax analysis is used to select candidate skills. The non-target domain set is used to calculate the domain relevance of each word in candidate skills to obtain candidate term domain relevance. This is then integrated into the C-value of candidate skills, sorted in descending order, and the top N candidate skills are selected as extracted skills.

Human evaluation determines whether the top N candidate skills are correct to calculate precision. Meanwhile, 500 recruitment texts are randomly selected from the target domain set, and skill information is manually identified to test recall. Finally, the F-measure is obtained by combining precision and recall to evaluate the method. Precision, recall, and F-measure are calculated using formulas (5)-(7):

$$\text{Precision} = \frac{\text{Number of correctly extracted skills}}{\text{Number of extracted skills}} \times 100\%$$

$$\text{Recall} = \frac{\text{Number of correctly extracted skills}}{\text{Number of actual skills}} \times 100\%$$

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%$$

4.3 Experimental Results

4.3.1 Effectiveness of Candidate Skill Selection Based on Dependency Syntax Analysis The experiment first verifies the effectiveness of candidate skill selection based on dependency syntax analysis. We compare the traditional candidate skill selection method with our dependency syntax analysis-based method, both using C-value to rank candidate skills. The former is called C-value, and the latter is called DepC-value. The comparison results are shown in Figures 5 [Figure 5: see original paper]-7 [Figure 7: see original paper]. As seen in Figures 5-7, DepC-value achieves higher precision, recall, and F-measure than C-value, demonstrating the effectiveness of using dependency syntax analysis for candidate term selection. The dependency syntax analysis-based candidate skill selection method addresses the characteristic that skill requirement statements in online recruitment texts are typically verb-object structures, filtering out unnecessary noisy verbs and thereby improving skill extraction precision and recall. Notably, this method also significantly reduces the number of candidate skills, shortening subsequent computation time.

Table 3 shows the top 10 noisy verbs filtered out by dependency syntax analysis. As shown in Table 3, dependency syntax analysis can accurately filter noisy verbs, thereby reducing incorrect candidate skills.

Table 4 lists the top 10 skills extracted by C-value and DepC-value methods, with verbs removed by dependency syntax analysis shown in bold. Table 4 shows that since the traditional candidate term method cannot filter unnecessary verbs, and candidate skills containing these verbs are abundant, this not only increases computation time but also results in high C-values, reducing skill extraction precision and recall. In contrast, the dependency syntax analysis-based candidate term selection method can effectively filter some noisy verbs, achieving higher skill extraction precision and recall.

4.3.2 Effectiveness of Domain Relevance Measurement Next, the experiment evaluates the effectiveness of domain relevance measurement. Building on the previous DepC-value method, we incorporate domain relevance with C-value to re-evaluate candidate skills, resulting in the DepDRC-value method. The comparison between DepC-value and DepDRC-value methods is shown in Figures 8 [Figure 8: see original paper]-10 [Figure 10: see original paper]. As seen in Figures 8-10, DepDRC-value achieves significantly higher precision, recall, and F-measure than DepC-value, indicating that incorporating domain relevance measurement of candidate skills can substantially improve skill extraction performance.

Table 5 shows the top 10 skills extracted by DepC-value and DepDRC-value methods, with correct skills shown in bold. Table 5 demonstrates that DepDRC-value effectively reduces the DRC-value of low domain-relevance candidate terms such as “相关专业” (related major) and “工作经验” (work experience), while increasing the DRC-value of high domain-relevance candidate skills like “SQL 语

句" (SQL statements) and "Linux 常用命令" (Linux common commands), thereby achieving higher skill extraction precision, recall, and F-measure.

4.3.3 Comparison with Other Methods To verify the effectiveness of the proposed method, we compare four approaches: (1) C-value: using traditional candidate skill selection and C-value measurement; (2) MIC-value: using traditional candidate selection and incorporating mutual information into C-value; (3) EnC-value: using traditional candidate selection and incorporating adjacency entropy into C-value; and (4) DepDRC-value: our proposed method using dependency syntax analysis for candidate selection and incorporating domain relevance.

The results are shown in Figures 11 [Figure 11: see original paper]-13 [Figure 13: see original paper]. As seen in Figures 11-13, DepDRC-value achieves significantly higher precision, recall, and F-measure than other methods, indicating that C-value, MIC-value, and EnC-value are not suitable for skill extraction from online recruitment texts. Our proposed DepDRC-value method, which leverages dependency syntax analysis and incorporates domain relevance information, substantially improves the precision, recall, and F-measure of C-value.

Table 6 lists the top 10 skills extracted by MIC-value, EnC-value, and DepDRC-value methods, with correct skills shown in bold. Table 6 shows that MIC-value uses mutual information to measure word cohesion in candidate skills. However, since some non-skill strings also appear frequently, these strings have high mutual information values, leading to incorrect extraction results. EnC-value uses adjacency entropy to measure the uncertainty of words adjacent to candidate terms. However, some non-skill candidate strings have rich adjacency information, such as "熟练使用," which can connect to many types of information and thus has high adjacency entropy, causing incorrect extraction results. DepDRC-value uses dependency syntax analysis and introduces auxiliary online recruitment text sets to effectively measure the domain relevance of candidate skills, overcoming the shortcomings of C-value and thereby improving skill extraction precision and recall.

Conclusion

Online recruitment information often contains specific descriptions of skill requirements for positions, reflecting current market demands for talent. Therefore, analyzing online recruitment information can help understand society's skill requirements for domain professionals. However, online recruitment information is typically unstructured text, and traditional skill requirement analysis methods usually require manual extraction of skills from recruitment texts. Manual skill extraction cannot meet the requirements of large-scale, unstructured online recruitment information analysis. This paper addresses the characteristics of online recruitment texts by using dependency syntax analysis to select candidate skills and proposing the concept of skill domain relevance, which is integrated into the C-value method to automatically extract skills from online recruitment

texts. Experiments demonstrate that the proposed method can automatically, quickly, and accurately extract skills from massive online recruitment texts. Future work will attempt to analyze skill requirements for popular positions based on extracted skill information, providing instructive position skill requirement information for students, teachers, and universities.

References

[1] WOWCZKO I. Skills and vacancy analysis with data mining techniques[J]. Informatics, 2015, 2(4): 31-49.
[2] KIM J Y, LEE C K. An empirical analysis of requirements for data scientists using online job postings[J]. International journal of software engineering and its application, 2016, 10(4): 161-172.
[3] MAURO A D, GRECO M, GRIMALDI M, et al. Beyond data scientists: a review of big data skills and job families[C]//Proceedings of the 2016 international forum on knowledge asset dynamics. Berlin: Springer International Publishing, 2016: 1844-1857.
[4] Lü Bin, ZHANG Tong, ZHOU Jue. General intelligence profession and intelligence professionals for organizations—analysis based on organizational recruitment web page information mining (Part 1)[J]. Library and Information Service, 2009, 53(4): 19-23.
[5] LI Guoqiu, SANG Peiming. Intelligence process—the core of intelligence profession: problem domain and methodology—analysis based on organizational recruitment web page information mining (Part 2)[J]. Library and Information Service, 2009, 53(4): 24-27.
[6] XIA Huosong, PAN Xiaoting. Research on the relationship between big data academic research and talent demand based on Python mining[J]. Journal of Information Resources Management, 2017, 7(1): 4-12.
[7] HUANG ?, WANG Kaifei, WANG Shanshan, et al. Investigation on recruitment requirements for data positions and implications for talent cultivation in library and information science[J]. Library and Information Knowledge, 2016, 6(1): 42-50.
[8] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms:. the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2): 115-130.
[9] SODHI M S, SON B G. Content analysis of OR job advertisements to infer required skills[J]. The journal of the Operational Research Society, 2010, 9(1): 1315-1327.
[10] ZHAO M, JAVED F, JACOB F, et al. lSKILL: a system for skill identification and normalization[C]//Proceedings of the twenty-seventh conference on innovative applications of artificial intelligence. Palo Alto: AAAI, 2015: 4012-4017.
[11] XU T, ZHU H, ZHU C, et al. Measuring the popularity of job skills in recruitment market: a multi-criteria approach[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. Menlo Park: AAAI, 2018: 3013-3028.

[12] ZHAN Chuan. Analysis of professional skill requirements based on text mining—taking e-commerce major as an example[J]. Library Tribune, 2017, 5(1): 116-123.
[13] XIA Lixin, CHU Lin, WANG Zhongyi, et al. Construction of employment knowledge demand relationship based on web text mining[J]. Library and Information Knowledge, 2016, 169(1): 94-100.
[14] BASTIAN M, HAYES M, VAUGHAN W, et al. LinkedIn skills: large-scale topic extraction and inference[C]//ACM conference on recommender systems. New York: ACM, 2014: 1-8.
[15] LIU Ruilun, YE Wenhao, GAO Ruiqing, et al. Research on text clustering based on big data position requirements[J]. Data Analysis and Knowledge Discovery, 2017, 12(12): 32-40.
[16] CONRADO M D, PARDO T A, REZENDE S O. A machine learning approach to automatic term extraction using a rich feature set[C]//The North American chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2013: 16-23.
[17] PIAO S, FORTH J, GACITUA R, et al. Evaluating tools for automatic concept extraction: a case study from the musicology domain[C]//Proceedings of digital features. Piscataway: IEEE, 2010: 78-85.
[18] SPASIC I, GREENWOOD M, PREEC A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 4(1): 27-42.
[19] MAYNARD D, ANANIADOU S. Identifying terms by their family and friends[C]//Proceeding of the 18th conference on computational linguistics. New York: ACM, 2000: 530-536.
[20] ZHOU Shuangshuang, XU Jin'an, CHEN Yufeng, et al. A microblog new word discovery method combining rules and statistics[J]. Computer Applications, 2017, 37(4): 1044-1050.
[21] ZHAO Jingsheng, ZHU Qiaoming, ZHOU Guodong, et al. A survey on automatic keyword extraction[J]. Journal of Software, 2017, 28(9): 2431-2449.
[22] LIU Huaijun, CHE Wanxiang, LIU Ting. Feature engineering for Chinese semantic role labeling[J]. Chinese Journal of Information Processing, 2007, 21(1): 79-84.
[23] Harbin Institute of Technology Language Technology Platform LTP[EB/OL].[2018-12-30]. http://ir.hit.edu.cn/demo/ltp.
[24] CHE W, LI Z, LIU T, et al. A Chinese language technology platform[C]//The 23rd international conference on computational linguistics. Stroudsburg: ACL, 2010: 3-16.

Author Contributions

Yu Yan: Proposed research ideas, designed research plan, conducted experiments, and wrote the paper;
Chen Lei: Data cleaning;
Jiang Jinde: Analyzed data and revised the paper;
Zhao Naixun: Revised the paper.