Investigating Natural Language Processing Techniques for a Recommendation System to Support Employers, Job Seekers and Educational Institutions

Koen Bothmer, Tim Schlippe

IU International University of Applied Sciences
tim.schlippe@iu.org

Abstract. Skills are the common ground between employers, job seekers and educational institutions which can be analyzed with the help of natural language processing (NLP) techniques. In this paper we explore a state-of-the-art pipeline that extracts, vectorizes, clusters, and compares skills to provide recommendations for all three parties—thereby bridging the gap between employers, job seekers and educational institutions. Our best system combines Sentence-BERT [1], UMAP [2], DBSCAN [3], and K-means clustering [4].

Keywords: AI in education, recommender system, recommendation system, up-skilling, natural language processing.

1 Introduction

There are often gaps between the skills that are needed in the labor market, the skills that job seekers[1] have and the skills that are taught in educational institutions [5]. Connecting and supporting all three players allows the greatest possible exchange of information and satisfies their needs. However, they usually use AI in isolation from one another [6,7,8,9]. Since skills are their common ground which can be analyzed with the help of AI, we investigate several NLP techniques to extract, vectorize, cluster and compare skills. Then we combine the optimal methods in a pipeline which serves as the basis for our application *Skill Scanner*[2] [10] that outputs statistics and recommendations about missing and covered skills for all three players. Our goal was to help employers, job seekers and educational institutions adapt to the job market's needs. Consequently, we used job postings, which represent the job market's needs, as reference. These representative skills, which we draw from a large set of job postings, are referred to as "*market skills*" in this paper. As companies hiring data scientists find that it is difficult to find a so-called "unicorn data scientist" [11], we conducted our experiments and analysis using companies' job postings for a data scientist position, job seekers' CVs for that position, and a curriculum from a master's program in data science. But our investigated methods can be applied to other job positions as well.

[1] "job seeker" refers to individuals who wish to apply for or advance in a job.
[2] https://github.com/KoenBothmer/SkillScanner

2 Related Work

Automatically ranking CVs is a valuable tool for employers. For example, [12] rank candidates for a job based on semantic matching of skills from LinkedIn profiles and skills from their job description, relying on a taxonomy of skills. Recent advancements in NLP offer opportunities to improve these methods: [6] use word embeddings from Word2Vec [13] to match CVs to jobs. [9] combine a knowledge graph and BERT for finding suitable candidates in a corpus of CVs. Recommendation systems for job seekers have been investigated by [14,15,16]. As in the systems for employers, text data from social media profiles such as LinkedIn or Facebook is usually processed [8,17]. [18] give a systematic review of recent publications on course recommendation. Most related work focuses on recommending courses to potential students. They report a growing popularity of data mining techniques. To cope with different levels of abstraction and synonyms in the course materials and students' documents, they first cluster the content, which they can then compare. K-means [4] is usually used for this.

3 NLP to Extract, Vectorize, Cluster and Compare Skills

For a certain job position, our pipeline (1) takes a CV, a job posting or a learning curriculum as input, (2) extracts the skills of the provided document, (3) compares the document's extracted skills to a skill set which represents the market's needs (*market skills*) and (4) returns information of which *market skills* are covered or missing in the document. Figure 1 visualizes the steps of our corresponding NLP pipeline.

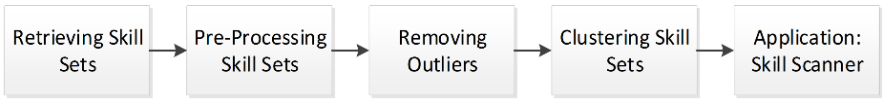

Fig. 1. Pipeline to Extract, Vectorize, Cluster, and Compare Skills.

3.1 Retrieving Skill Sets: Extract Skill Requirements

In job postings, CVs and learning curricula, skills are usually expressed in bullet points. Therefore, we developed keyword- and rule-based techniques to extract bullet points from these sources. Furthermore. we used the *BeautifulSoup* package to gather and extract 21.5k bullet points from 2,633 job postings for data scientists in English from Indeed.com and Kaggle.com which represents the market's needs (*market skills*). Since some bullet points in a job posting are not skill requirements, we analyzed methods to deal with outliers that are not skill requirements as described in Section 3.4.

3.2 Vectorizing Skills: Map Skill Requirements to Semantic Vector Space

To compute distances between skills, we mapped the skills to a semantic vector space. To represent the skills which usually consist of several words, we investigated stacking and averaging word embeddings in a skill which were produced with Word2Vec [13]

and GloVe [19]. In addition, we explored sentence embeddings. Sentence-BERT (44.2%) [1], a modification of the BERT transformers, outperformed word embeddings like GloVe (39.5%) by 12% in Silhouette score [20] at the end of our pipeline.

3.3 Removing Outliers from Skill Requirements

To remove outliers in the vectorized skills and allow our clustering techniques to perform better, we reduce the dimensionality of the feature space created by Sentence-BERT. For that we experimented with combinations of PCA [21], UMAP [2], and DBSCAN [3]. Using UMAP to reduce the vectorized skills to two dimensions and DBSCAN to remove outliers in the 2-dimensional (2D) space performed best according to our manual checks and reduced the 21.5k potential skills retrieved with our web scraper to 18.8k skills. However, since the 2D vectors did not contain enough information for further analysis of the skill set, we applied another clustering to the original 768-dimensional vectors that remained after removing outliers.

3.4 Clustering Skills

To find comparable skills despite different levels of abstraction and synonyms in job postings, CVs and learning curricula, we use a clustering approach. The benefit of our clustering approach compared to a taxonomy is that our model can pick up new skills without the need to update a taxonomy. K-means clustering has been successfully used in clustering word embeddings [22] and is adaptable and scalable [4]. Consequently, we used K-means to cluster our 768-dimensional vectors with the cosine distance as the distance metric. K was chosen as 31 with the highest Silhouette score of 44%.

3.5 Skill Scanner: Comparison and Analysis

After retrieving clusters and vectors representing the skill of each cluster, we perform mathematical operations to find covered and missing skills regarding the job market's demand which are then visualized in reports for employers, job seekers, and educational institutions. More information on the visualization of our reports is given in [10].

4 Conclusion and Future Work

The labor market dictates what job seekers should learn, and educational institutions should teach. Therefore, our system processes skills in job postings, CVs, and curricula and outputs recommendations for employers, job seekers, and educational institutions based on present and missing skills and their importance to employers. With our clustering approach we do not have to update a taxonomy as skill requirements change. Future work may be to apply our pipeline to other job positions and expand it to other domains. Furthermore, as we used the pre-trained Sentence-BERT it may be analyzed if a fine-tuned Sentence-BERT leads to further improvement.

References

1. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP-IJCNLP (2019)
2. McInnes, L., Healy J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv, abs/1802.03426 (2018)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD. AAAI Press, 226–231 (1996)
4. Lloyd, S.P.: Least Squares Quantization in PCM. Techn. Report RR-5497, Bell Lab (1957)
5. Palmer, R.: Jobs and Skills Mismatch in the Informal Economy. 978-92-2-131613-8 (2017)
6. Fernández-Reyes, F.C., Shinde, S.: CV Retrieval System Based on Job Description Matching Using Hybrid Word Embeddings, Computer Speech & Language, vol 56 (2019)
7. Geyik, S.C., Guo, Q., Hu, B., Ozcaglar, C., Thakkar, K., Wu, X., Kenthapadi, K.: Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned. SIGIR (2018)
8. Guruge, D.B., Kadel, R., Halder, S.J.: The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research Data, 6(2), 18 (2021)
9. Wang, Y., Allouache, Y., Joubert, C.: Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT. DBKDA (2021)
10. Bothmer, K., Schlippe, T.: Skill Scanner: Connecting and Supporting Employers, Job Seekers and Educational Institutions with an AI-based Recommendation System. The Learning Ideas Conference 2022 (15th annual conference), New York, New York (2022)
11. Baškarada, S., Koronios, A.: Unicorn Data Scientist: The Rarest of Breeds, Program: Electronic Library and Information Systems, Vol. 51 No. 1, pp. 65–74. (2017)
12. Faliagka, E., Iliadis, L. Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A. Tzimas, G.: On-line Consistent Ranking on E-Recruitment: Seeking the Truth Behind a Well-Formed CV. Artif Intell Rev 42, pp. 515–528 (2014)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ICLR (Workshop Poster) (2013)
14. Si-ting, Z., Wenxing, H., Ning, Z., Fan, Yang: Job Recommender Systems: A Survey. ICCSE (2012)
15. Hong, W., Zheng, S., Wang, H., & Shi, J. (2013). A Job Recommender System Based on User Clustering. J. Comput., 8, 1960-1967.
16. Alotaibi, S: A Survey of Job Recommender Systems. Int. J. Phys. Sci. (2012)
17. Diaby, M., Viennet, E., Launay, T.: Toward the Next Generation of Recruitment Tools: An Online Social Network-Based Job Recommender System. ASONAM (2013)
18. Deepani B. Guruge, Rajan Kadel, and Sharly J. Halder: The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research. Data 6, no. 2: 18. (2021)
19. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. EMNLP (2014)
20. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 20: 53–65. (1987)
21. Pearson, K.: On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 2 (11): 559–572 (1901)
22. Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G.; Does Deep Learning Help Topic Extraction? A Kernel K-Means Clustering Method with Word Embedding. Journal of Informetrics, 12 (4), 1099–1117 (2018)