Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT

Yan Wang, Yacine Allouache, Christian Joubert

Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT

Yan Wang
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: yan.wang2@altran.com

Yacine Allouache
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: yacine.allouache@altran.com

Christian Joubert
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: christian.joubert@altran.com

***Abstract*—Recruiter and candidate are the two main roles in the process of employment. Even though there is an abundance of job openings and a scarcity of qualified candidates to fill those openings, the objective is to offer only the profiles that fit the requirements of clients. Bidirectional Encoder Representations from Transformers (BERT) have been proposed in 2018 to better understand client searches. The challenges today are the frequent evolution of the experiences in Curriculum Vitae (CV) and the need of adaptive data for the specific staffing tasks of BERT. In this paper, we present an approach of ranking candidates based on competence keywords. There are four stages. First, we use Term Frequency–Inverse Document Frequency (TF-IDF) Vectorizer to calculate the score of matching between a competence keyword and a corpus of CVs. Second, we apply the Weighted Average Method to calculate a global score of CV based on two types of competence keywords – function and specialty. Third, we construct a Knowledge Graph (KG) from the structured Competence Map (CMAP), which can classify the relationships of bidirectional association and aggregation. At last, we propose to use the Named-Entity Recognition (NER) and Masked Language Modeling (MLM) of BERT to better identify tokens from the input inquiries of the client. The experiments are using the CVs from the HR (Human Resource) management system of Altran.**

Keywords—CV; CMAP; Knowledge Graph; BERT; TF-IDF Vectorizer; NER; MLM; HR-analytics; Job Matching.

I. INTRODUCTION

HR-analytics is always essential for companies to improve workforce performance. It allows many companies to exploit the immense amounts of data collected from their customers, their markets, social networks, real-time applications, and even the cloud. There is a coherent connection between engagement, performance and profit. It is imperative to generate performance results at all levels of the organization in order to take a position in the market and to stimulate growth. Recruiting talented candidates is not enough. However, it is important that people are assigned to specific roles where their talents will have the greatest impact on achieving company goals and where they are most likely to remain fully engaged. Job matching involves defining superior performance of each position and using objective criteria to determine who gets employed. Traditional hiring methods that only use a job description and a list of desirable candidates, technical, educational and professional experiences as filters, with a favorable interview are not working effectively. The process of job matching goes beyond conventional employment methods to create the most comprehensive definition of the job. This makes clients choose the right person for the job that suits them best. The result is a person who is happier at work and who makes good progress in meeting performance goals. Job Matching issues always exist in the center of operational and staffing concerns.

The similarity calculation is the mechanism that consists in evaluating the distance between two objects. Similarity measurements make it possible to solve problems from various fields such as text mining, image recognition, Nature Language Processing (NLP), computer vision, speech processing, image processing, and so on. In recent years, the recruiting process, Chatbot, search engines and recommendation systems have brought these technologies up to date. In this paper, we focus on the search engine for recruitment and staffing whereas Chatbot and recommendation systems are studied in future work.

Most staffing search engines do not consider about the evolution of the experiences. The objective of this paper is to use the NER of BERT [1] to extract relevant competence keywords from candidate experiences and client requests. To achieve this, the model of BERT that we use is fine-tuned by the vocabulary of competence keywords using the MLM of BERT. A Knowledge Graph is proposed for a semantic structure to help users find more precise results. However, the vector BERT assigns to a word is a function of the entire sentence. The vectors can be different with the same word. To deal with this problem, TF-IDF Vectorizer and Weighted Average Method are proposed to calculate the score based on the static vector.

The paper is organized as follows. Section 2 provides the related work of information parser, job matching and BERT. Section 3 explains the approach of extracting competence keywords and constructing the KG for recommendations. Section 4 shows the experiment of comparing the tool BERT with Linx. The conclusion and perspective are in Section 5.

II. RELATED WORK

Most of staffing software for recruiting and staffing in information extraction from CVs is either privatized by companies as an in-house tool or sold as a commercial

 ISBN: 978-1-61208-857-0

product, for instance the private HR management system Linx in Altran. Artificial intelligence has contributed to great success in this field, but due to the confidential rule, some works are not well seen by the research community. Resume parsers have been proposed to extract information from the Internet [2], Github [3] and PDF [4], as well as from the general non-LinkedIn formats [5]. In case of Linx, only pre-processing is required instead of the parser.

For the purpose of improving the matching rate between jobseekers and available jobs, an ontology based expert system [6] can improve the accuracy of this matching. [7] proposes to extend the matching to multiple slots available to accept contracts. [8] presents a survey of exact string matching and approximate string-matching algorithms. Machine learning can continue to improve the performance of matching rate such as unsupervised feature extraction [9], using Convolutional Neural Network (CNN) [10] and deep Siamese network [11]. At last, by combing the knowledge graph and recommendation system, we expect to improve the matching quality with the help of the relationships between entities, for example, by using embedding-based methods [12] or path-based methods [13]. Considering the user preference [14] can also be interesting. In this paper, we consider the matching based on the competence keywords.

BERT is known as a more powerful and efficient technique than the other NLP tools like RNN, CNN and LSTM, which understands inquiries better than ever before. The embedding models word2vec [15] and GloVe [16] have been presented to be less effective in documents recognitions. The performance also differs from section to section [17]. In recent years, BERT has been used for sentence classification based on the suitability of job description [5] and semantic search over the corpus [18]. BERT will become more and more important in the staffing software.

III. THE PROPOSED APPROACH

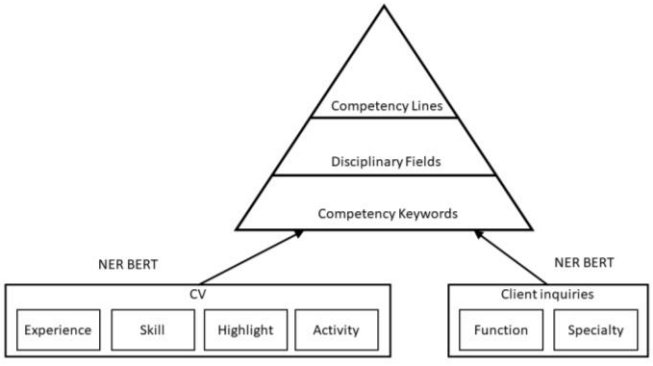

Figure 1. Structure of proposed approach to construct the KG.

The structure of the proposed approach is shown in Figure 1. In Section III.A, we define the competency keywords as the vocabulary. Section III.B and III.C will explain how to calculate the matching score. The pyramid represents the structure of the knowledge graph in Section III.D. Section III.E proposes to use MLM and NER of BERT to identify the features in client inquiries. The process of this approach is first to extract the keywords using NER of BERT, and then return a list of the ranked candidates from Knowledge Graph.

A. Competency keyword

The structure of a CV from Altran contains several sections. In this paper, we focus on four sections – experience, core skills, skill keywords and activity keywords. Competency keywords are used as a vocabulary for a client to search for candidates. Competency keywords are collected from the core skills, skill keywords and activity keywords sections, where they are stored by the HR system of Altran. But they are not standardized and refined. Accordingly, some keywords may have the same meaning in the semantic competency. In order to deal with the confusing keywords, a pre-processing of data is required. The purpose of cleaning is to remove ambiguous data, for example "JS" represents "Java Script". The competency keywords are also classified into two types – function and specialty. Function refers to the position of the job coming from activity keywords such as "software developer" and specialty refers to the specific skill coming from skill keywords such as "mobile application".

B. TF-IDF Vectorizer for Matching Score

The TF-IDF method is used at first to represent text as a vector of dimension D such that D is the number of words in the vocabulary - competency keywords. So, for each word of the vocabulary, we compute its TF-IDF in each section of the document – experience, core skills, skill keywords and activity keywords. To compute this score, we proceed in two steps. The first step is to compute the TF, which is the number of occurrences of the word in the document. Then, we calculate the IDF, which is a metric of the importance of the word. The intuition behind the definition of IDF is if a term is present in more documents, then it is more important.

$$w_{i,j} = \mathrm{t}f_{i,j} \times \mathrm{idf_i} \quad (1)$$

$$\mathrm{id}f_i = \log\left(\frac{1+N}{1+df_i}\right) \quad (2)$$

$\mathrm{w}_{i,j}$ = score of term i in document j
$\mathrm{t}f_{i,j}$ = number of occurrences of term i in document j
$\mathrm{d}f_i$ = number of documents containing the term i
N = total number of documents

TF-IDF Vectorizer allows us to have an explainable representation, as each dimension *i* of our vector represents the score of word *i* in a document *j* or in the client query.

C. Weighted Average Method for Global Score

The text of each section in a CV – experience, core skills, skill keywords and activity keywords can be given a matching score. However, the strength of the connection in each of the four sections may be different. A Weighted Average Method has been proposed to calculate a global score concerning all four sections. The weighted average method is defined by the following equation:

$$\mathrm{S} = \frac{\sum \mu(s) \cdot s}{\sum \mu(s)} \quad (3)$$

where s is the score of matching in each section and μ represents the possibility of each section. The possibility of each section is pre-defined, but the value may change according to the types of competence keywords. For example, if the competence keywords are in the type of function like “software developer”, the possibility of activity keywords is higher than core skills and skill keywords. Conversely, the same is true for the type of specialty like “mobile application”.

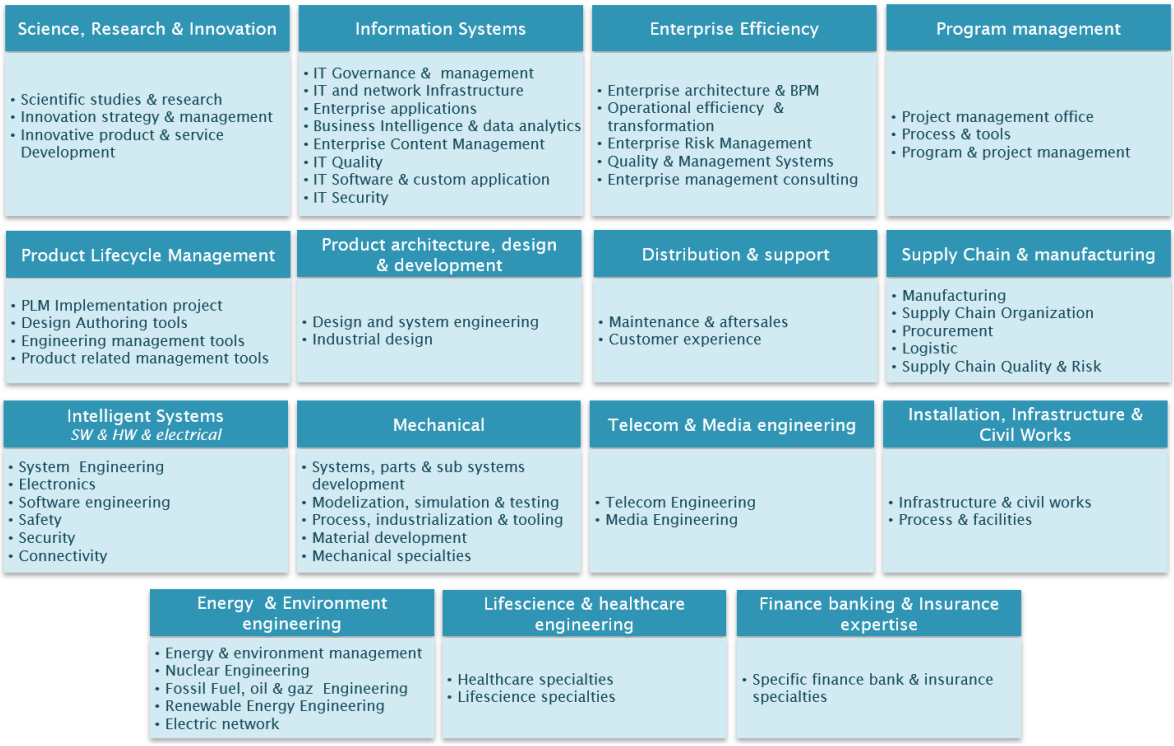

Figure 2. The competency lines and disciplinary fields in candidate environment of CMAP

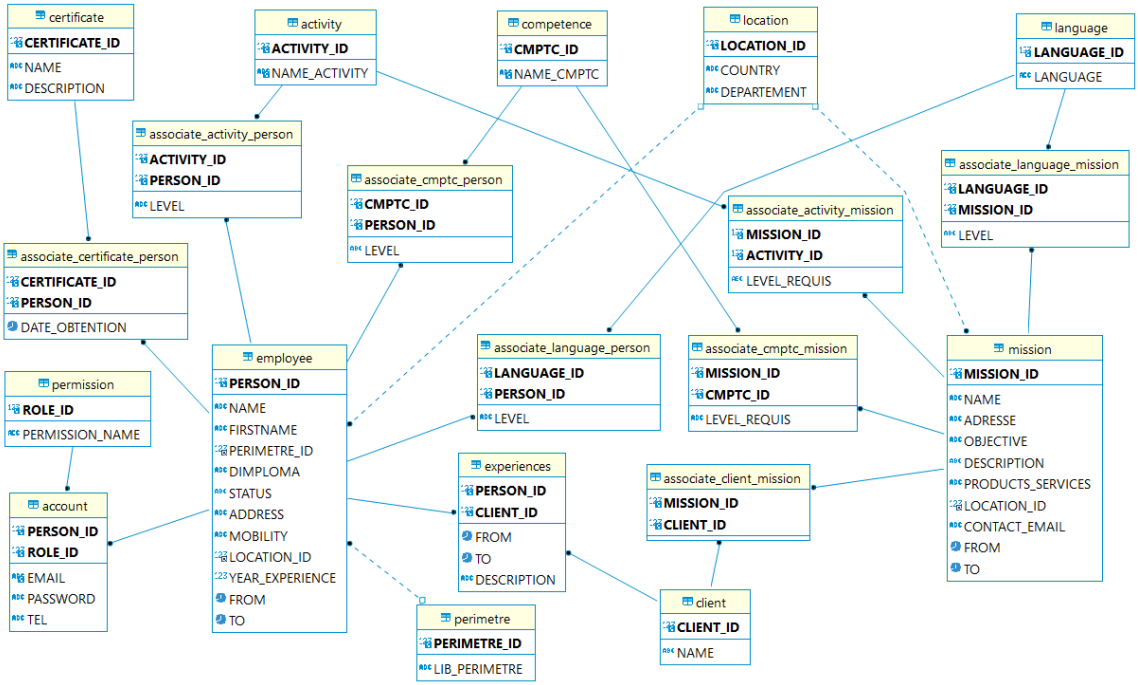

Figure 3. ER diagram of relational database for the profil of candidate and mission.

D. Knowledge Graph from CMAP

CMAP has been proposed as a homogeneous description of the competencies in Altran. The purpose is to allow and support competency-based management. It is not only a knowledge management system, but also a strategic workforce planning. CMAP contains a client environment and a candidate environment. Figure presents the structure of the candidate environment. The keywords in dark blue belong to the competency lines and the keywords in light blue belong to the disciplinary fields.

In this paper, a knowledge graph is proposed to ensure search results are contextually relevant to requirements. The first version of the KG with structed data is stored in the graph database [17]. The Neo4j Graph Platform that we choose for KG is an example of a tightly integrated graph database and algorithm-centric processing, optimized for graphs. There are three labels in the KG, where the label marks the node as the part of a group. The first label is constructed by the competency lines and the second label is generated from the disciplinary fields. The relationship of aggregation is identified such that the first label contains the second label based on CMAP and each second label contains several competency keywords represented by the third label. In addition, two competency keywords can have a relationship of bidirectional association which indicates a close connection between each other. The first version of the labels in the Knowledge Graph is based on structured data. It is a tree and sparse graph as well as bipartite graph distinguished between function and specialty.

E. MLM and NER using BERT

BERT [1] is a feature extractor based on deep Neural Probabilistic Language Models proposed in 2018. MLM is a fill-in-the-blank task, where a model uses the whole context words to predict the masked word. In our approach, BERT is first trained by the MLM method to modify the model distribution to be specific to our domain. This method consists of masking the tokens of a sequence with a masking token <**MASK**> and asking the model to fill this mask with the appropriate token. These tokens come from the keywords of Knowledge Graph, such that if the word belongs to the Knowledge Graph, we mask it with a probability $\boldsymbol{P}$. If not, we mask it with a probability $\boldsymbol{1-P}$. A threshold of $\boldsymbol{P > 0.5}$ is set to allow the model to focus on the competency keywords. This allows the model to be attentive to both the right context - tokens on the right of the mask, and the left context - tokens on the left of the mask.

Secondly, we continue the training of BERT with the NER method in order to extract the competency keywords from the profiles of candidates and the requests of clients. NER can classify tokens based on a class, for example, identifying a token as a person, an organization, or a location. In this approach, we use this technique to classify tokens as FUNC, SEP or NONE, such that *FUNC* means function, *SEP* means specialty, and *NONE* means a simple token. Therefore, we can further enrich our Knowledge Graph with the new competency keywords detected by BERT, which allows us to create a cycle of both refining BERT with the Knowledge Graph and developing the Knowledge Graph with BERT.

IV. EXPERIMENT

A. Dataset

In the HR management system of Altran - Linx, there is a searching engine for CVs. Location, keyword, availability, industry sector, competence domain, activity, certificates and years of experience are the options for searching. In order to evaluate our proposed approach, we make a search of "engineer" and "available" in "France". In total, 106 CVs are exported in Excel, and 45 competence keywords are collected after pre-processing of removing confusing keywords. The interesting part of these CVs are filtered and then stored in the relational database MySQL shown in Figure 3. In this Entity Relationship (ER) diagram, the table employee stores the main information of the candidate as well as the availability time and years of experiences. For privacy, we keep name and personal contact information of the candidate as empty. Another table stores the content of experience. Five key tables store the keywords of certificate, function, specialty, location and language, respectively. At last, the table of permission is used to define the authentication of the user to login the system. The tool BERT implemented by the approach proposed in this paper is executed as a Micro-Service connected with the frontend.

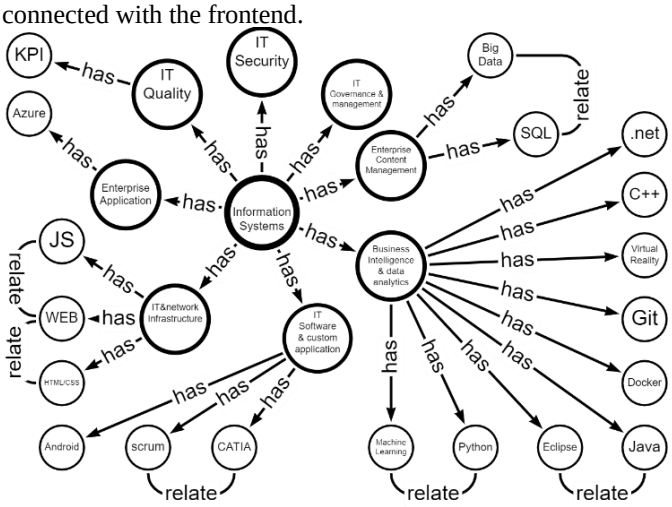

Figure 4. Knowledge Graph in the label of Information System.

The Knowledge Graph is constructed in Neo4j using 45 competency keywords and CMAP. To be different, MySQL is used to store the CVs and other original data and Neo4j is used to construct the KG for recommendations of competence keywords. Figure 4 presents the Knowledge Graph in the label of Information System. All the eight disciplinary fields surrounding this label have the aggregation relationship of "has", for example "Information System" has a relation of "has" with "IT Quality", "IT Security", etc. 19 competency keywords are attached to the corresponding disciplinary field. Among the competency keywords, some of them can have a

bidirectional association relationship of "relate", for example "Java" is related to "Eclipse". The relationship of "relate" can recommend the related keyword with explanation based on the request of the client.

B. Results

The CVs extracted from Linx are written in both English and French language. A multi-language tokenizer is used for pre-processing at first. To perform a search using BERT, we extract the competency keywords from the profiles of candidates, which allows us to represent a candidate as a list of competency keywords. The profile is represented by ID and the competence keyword is represented by KEY. Then, we apply the TF-IDF model to calculate the scores of KEY in each section of the profile and then Weighted Average Method to calculate the global stores of KEY. The global scores are stored by the index of ID:KEY and the inverted index KEY:ID in a Redis database. Secondly, for the query of clients, we have as input a list of words that represent the criteria searched by the client. This query is passed to the NER tool, especially multi-language DistilBERT base model [20], to extract the competency keywords. The model of DistilBERT that we choose can understand both English and French language. Since the inverted index contains KEY:ID, Redis can return a list of profiles containing these competency keywords sorted in decreasing order.

The NER tool is implemented using the model of DistilBERT-base-multilingual-case and is pre-trained by MLM. This tool is able to understand the language of both English and French. The objective of this tool is to use BERT methods to automate the recognition and extraction of keywords from the context.

At last, we compare the results obtained with Linx and BERT. However, Linx is a tool where the search principle is only based on the competency keywords presented in the platform and we are limited on how to express our query such that it must be written in a Boolean way containing only the competency keywords of the skills that we want. For this reason, we have implemented this tool using BERT in order to express our needs with natural language. So, for our query "Java" we chose to search the first 3 consultants that are proposed by our models. With the limitation of the pages in this paper, we choose three top ranking candidate results from both Linx and BERT in Table 1. The sections Experience, Core Skills and Skill keywords contain original data that we use for analysis. It is obvious to see that Linx does not pay attention to the experiences of candidates and the results of Linx only focuses on the sections of core skills and skill keywords, while BERT considers both the experiences and the skills. All the three candidates from BERT have experiences of Java and this skill is also highlighted in the sections of core skills and skill keywords. Especially, the first candidate from BERT has the experience of "Eclipse" which is related to "Java" in the knowledge graph. The volume of CVs and the latency are not considered in this experiment as we focus on the measurement of semantic similarity.

TABLE I. THREE TOP RANKING CANDIDATE RESUTLS IN BOTH LINX AND BERT

Linx				**BERT**			
ID_ candidate	***Experience***	***Core Skills***	***Skill Keywords***	***ID_ candidate***	***Experience***	***Core Skills***	***Skill Keywords***
16874	Developer at ENGIE - France Developer within the Genesys team then WattsOn at GEM IS Consultant at SFR - France Project manager / IT Engineer MOE - Scalable and corrective maintenance of various applications. IS Consultant at SOCIETE GENERALE - France IT MOE Engineer - Scalable and corrective maintenance of several applications	Test automation, Analysis & development of software requirements, Software design	DDD, Microsoft Visual Studio 2010, Entity Framework 4.0, C# 4.0, Java TDD	346575	Developer at AIRBUS - France: Creation of a Platform for Icing studies Developer at AIRBUS - France: Prerevolution Software Developer at DASSAULT AVIATION - France: Eclipse RCP based Verification Tool development for SCADE ENSO	Modelling, Model-Based Systems Engineering, IT Test & Validation, Automation	JUNIT, Maven, EMF, Tortoise Git Java 8
17071	Study engineer at BNP PARIBAS - France Universal Plug application - migration from Exadata to AIX, optimization Technical consultant at SOCIETE GENERALE - France I2R - Performance optimization of the Oracle Exadata database; migration from Oracle 11gR2 to 12c Technical consultant at SOCIETE GENERALE - France Optimization of the AGORA-AIR application database	DBA study, Database development, System and database	Oracle PL/SQL, Oracle Exadata, Oracle 12c, Oracle SQL Developer, Java 8	67747	Developer at ALTRAN - France: Clinuikali project - Java Application development Developer at ALTRAN - France: Python Application development Testing & Validation Engineer at ALTRAN - France: ACS - Automatic Testing within Continuous Integration	Software design, Marketing studies and strategy, Integration Validation Verification & Qualification,	Python, Software Development, Java, CSS 3, Html5
15868	Tester at HP ENTERPRISE SERVICES COMMUNICATION & MEDIA SOLUTIONS - France System Tests and Functional Tests on a virtualized system of the 4G network	Functional testing and validation, Test and technical validation,	Collabnet Svn, Teamforge Svn, Microsoft Office,	67711	Developer at ACS - France: OA: The system quantifies the fatigue at work for an employee during his working hours. Developer at ACS - France:	Application WEB, Core network mobile circuits,	HTML, JavaScript, AngularJS, Angular, Java

Linx				BERT			
ID_ candidate	***Experience***	***Core Skills***	***Skill Keywords***	***ID_ candidate***	***Experience***	***Core Skills***	***Skill Keywords***
	core, 2G / 3G environment (MAP, Diameter, AAA, EIR protocols) Integrator at HP ENTERPRISE SERVICES COMMUNICATION & MEDIA SOLUTIONS - France Integration of software solutions - Pre-Integration Tests Management of off-shore teams Industrialization & production engineer at TOTAL - France Outsourcing control on data transfer applications Definition of new flows; Dedicated projects with high business impact	Collaboration and networking	HP ALM, Collaborative Tools		LBS: Tracking the movement of physical assets on indoor and outdoor topology, by scanning barcode labels attached to assets or using smart labels, such as LORA or Antiote labels, which broadcast their location. Developer at ACS - France: OA: The system quantifies the fatigue at work for an employee during his working hours.	Product design and development	

V. CONCLUSION

This paper presents a keyword-based search engine for recruitment and staffing using knowledge graph and BERT. NER of BERT pre-trained by MLM can better recognize the competence keywords from the corpus of CVs and the natural language of client inquires. The knowledge graph composed of CMAP and competency keywords can recommend good results in the neighborhood domain. The proposed approach provides a way to use BERT for this specific task. The experiment based on BERT shows a better performance on finding good candidates than Linx.

The future work contains three parts. A method of using BERT to compute the score of the word embedding is required to replace the TF-IDF Vectorizer; the first version of KG that we propose in this paper is based on structed data. A richer KG is needed with the properties of each node and weighted relation for the dynamic management in case of a new competence; in Figure 2, the information about mission is also stored in the relational databases. A recommendation system with KG embedding method is needed for the matching between the profile of a candidate and the description of a mission. Unifying knowledge graph learning and recommendation is highly suggested to improve the matching efficiency.

ACKNOWLEDGMENT

This paper is sponsored by Direction of Research and Innovation (DRI) of Capgemini Engineering in France.

REFERENCES

[1] 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.

[2] C. Mang, Online job search and matching quality. Ifo Working Paper, 2012.

[3] C. Hauff and G. Gousios, Matching GitHub developer profiles to job advertisements. IEEE/ACM 12th Working Conference on Mining Software Repositories. IEEE, pp. 362-366, 2015.

[4] J. Chen, L. Gao, and Z. Tang, Information extraction from resume documents in pdf format. Electronic Imaging, pp. 1-8, 2016.

[5] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. arXiv preprint arXiv:1910.03089, 2019.

[6] V. Senthil Kumaran and A. Sankar, Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT). International Journal of Metadata, Semantics and Ontologies, pp. 56-64, 2013.

[7] S. D. Kominers and T. Sönmez, Matching with slot‐specific priorities: Theory. Theoretical Economics, pp. 683-710, 2016.

[8] S. I. Hakak et al., Exact string matching algorithms: Survey, issues, and future research directions. IEEE Access, pp. 69614-69637, 2019.

[9] Y. Lin, H. Lei, P. C. Addo, and X. Li, Machine learned resume-job matching solution. arXiv preprint arXiv:1607.07657, 2016.

[10] C. Zhu et al., Person-job fit: Adapting the right talent for the right job with joint representation learning. ACM Transactions on Management Information Systems (TMIS), pp. 1-17, 2018.

[11] S. Maheshwary and H. Misra, Matching resumes to jobs via deep siamese network. Companion Proceedings of the The Web Conference 2018. pp. 87-88, 2018.

[12] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Y. Ma, Collaborative knowledge base embedding for recommender systems. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 353-362, 2016.

[13] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, Meta-graph based recommendation fusion over heterogeneous information networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 635-644, 2017.

[14] Y. Cao, X. Wang, X. He, Z. Hu, and T. S. Chua, Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. The world wide web conference. pp. 151-161, 2019.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[16] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532-1543, 2014.

[17] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, PROSPECT: a system for screening candidates for recruitment. Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 659-668, 2010.

[18] A. A. Deshmukh and U. Sethi, IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. arXiv preprint arXiv:2007.12603, 2020.

[19] M. Needham and A. E. Hodler, Graph Algorithms: Practical Examples in Apache Spark and Neo4j. O'Reilly Media, 2019.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

 ISBN: 978-1-61208-857-0