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

DBKDA 2021

Yan WANG, Yacine ALLOUACHE, Christian JOUBERT
Capgemini Engineering, Direction of Research and Innovation (DRI)
yan.wang2@altran.com, yacine.allouache@altran.com, christian.joubert@altran.com

Presenter: Yan Wang
Yan WANG

Dr. Yan Wang is a research project manager in the Direction of Research & Innovation (DRI) of Capgemini Engineering, France. He received his Master (2014) and PhD (2018) degree in Enterprise Engineering from University of Bordeaux. He received his Bachelor (2012) of Computer Science and Master (2014) of Software Engineering from Harbin Institute of Technology. He made two postdocs of Computer Science from Paris 1 Panthéon-Sorbonne University and Sorbonne University. His research interests include Discrete Event Modeling and Simulation, Process Mining, Fuzzy Logic, Constraint Programming, Reinforcement Learning. His current research is on NLP, Knowledge Graph, Recommender System.
Research and Development Project TNT

TNT (Talent Needs Trends) is a research and development project of the program Future of Engineering. The objective of this project is to propose a competence management system advanced and adapted to Capgemini. The purpose is to improve the synergy between skills, resources and customers, simplify the process of HR-Analytics.

TNT is launched from 2014 with a lot of propositions and development tools. In this paper, we focus on the proposal of the competence search engine using BERT and Knowledge Graph, with the purpose of improving the existing HR management tool Linx.
Problem of Competence Search Engine

- The rapid evolution of the competences of the candidates
- The continuing evolution of candidate experience and technology
- The rapid evolution of the requirements of the clients
- HR adaptation and treatment time is getting longer and longer
- The need of advanced Natural Language Processing (NLP) tool to understand the context of candidate competences and clients requirements
Challenge of BERT

Bidirectional Encoder Representations from Transformers (BERT) have been proposed to better understand client searches, a feature extractor based on deep Neural Probabilistic Language Model (J. Devlin et al. 2018).

- **Bidirectional**: A so-called bidirectional NLP model is a model that reads editorial content in its entirety and in both directions.
- **Masked Language Modeling (MLM)**: MLM is a fill-in-the-blank task, where a model uses the whole context words to predict the masked word.
- **Named-Entity Recognition (NER)**: NER can classify tokens based on a class, for example, identifying a token as a person, an organization, or a location.
- **Encoder**: The encoder is a neural network used to transform the input sequence into a vector representation of the sequence.
- **Decoder**: Second network after the encoder which is used to generate words sequentially.

The challenges today is the need of adaptive data for the specific staffing tasks of BERT.
Knowledge Graph (KG) has been proposed to discover the relation information with the property of powerful language understanding and rapid data analysis. It is first proposed in 2012 by Google, a theory of semantic structure combining applied mathematics, computer graphics, information visualization and machine learning. With the help of KG, users can get a more accurate recommendation as well as the explanations for recommended items. (Q. Guo et al. 2020)

The challenge today is the design and construction of Knowledge Graph based on competence keywords.
Example of CVs from Linx

The CVs that we use in this paper are from Linx, they can be exported as Excel files with the section of experience, core skills, activities, keywords, languages, diplomas, certifications, capacity available, location.

<table>
<thead>
<tr>
<th>CORPORATE ID</th>
<th>EXPERIENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>336048</td>
<td>Chef de projet chez ALTRAN - France: 01/2019 à Aujourd'hui (27 mois)</td>
</tr>
<tr>
<td></td>
<td>Road traffic management Consultant(e) chez UNIVERSIDAD DE LOS ANDES / MEALS DE COLOMBIA - Colombie: 03/2014 à 05/2014 (3 mois)</td>
</tr>
<tr>
<td></td>
<td>Analyse et amélioration de la chaîne logistique de MEALS de Colombie (Ice cream company in Colombia)</td>
</tr>
<tr>
<td></td>
<td>Analyste chez DOWER PEOPLE - Colombie: 09/2013 à 11/2013 (3 mois)</td>
</tr>
<tr>
<td></td>
<td>Analyse des employés compensation</td>
</tr>
</tbody>
</table>

**CORE SKILLS 0**
- Technologie pour la chaîne logistique
- Gestion de la planification

**CORE SKILLS 1**
- Gestion de l'approvisionnement

**CORE SKILLS 2**
- (Recherche, Innovation & Veille (Matlab))
- (Object-Oriented Programming)
- (Espagnol: Langue Maternelle)
- An 2019: logistique Université de Tec 2013: Production management

**CERTIFICATIONS 0**
- Industrial engineering-Universidad de los Andes-Colombia

**CERTIFICATIONS 1**
- UNB (INT-I)

**CAPACITY AVAILABLE**
- 100

**LOCATION**
- France

**WORKSITE**
- France - Vélizy (Topaz)
The proposed global database is used to save the input of CVs. It has five typical tables – certificate, activity, competence, location, language.
Competence Map (CMAP)

CMAP is a homogeneous description of the competencies across Altran.

<table>
<thead>
<tr>
<th>Science, Research &amp; Innovation</th>
<th>Information Systems</th>
<th>Enterprise Efficiency</th>
<th>Program management</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Scientific studies &amp; research</td>
<td>• IT Governance &amp; management</td>
<td>• Enterprise architecture &amp; BPM</td>
<td>• Project management office</td>
</tr>
<tr>
<td>• Innovation strategy &amp; management</td>
<td>• IT and network infrastructure</td>
<td>• Operational efficiency &amp; transformation</td>
<td>• Process &amp; tools</td>
</tr>
<tr>
<td>• Innovative product &amp; service development</td>
<td>• Enterprise applications</td>
<td>• Enterprise Risk Management</td>
<td>• Program &amp; project management</td>
</tr>
<tr>
<td></td>
<td>• Business Intelligence &amp; data analytics</td>
<td>• Quality &amp; Management Systems</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Enterprise Content Management</td>
<td>• Enterprise management consulting</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• IT Quality</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• IT Security</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• IT Software &amp; custom application</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• IT Security</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product Lifecycle Management</th>
<th>Product architecture, design &amp; development</th>
<th>Distribution &amp; support</th>
<th>Supply Chain &amp; manufacturing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• PLM Implementation project</td>
<td>• Design and system engineering</td>
<td>• Maintenance &amp; aftersales</td>
<td>• Manufacturing</td>
</tr>
<tr>
<td>• Design Authoring tools</td>
<td>• Industrial design</td>
<td>• Customer experience</td>
<td>• Supply Chain Organization</td>
</tr>
<tr>
<td>• Engineering management tools</td>
<td></td>
<td></td>
<td>• Procurement</td>
</tr>
<tr>
<td>• Product related management tools</td>
<td></td>
<td></td>
<td>• Logistic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Supply Chain Quality &amp; Risk</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intelligent Systems SW &amp; HW &amp; electrical</th>
<th>Mechanical</th>
<th>Telecom &amp; Media engineering</th>
<th>Installation, Infrastructure &amp; Civil Works</th>
</tr>
</thead>
<tbody>
<tr>
<td>• System Engineering</td>
<td>• Systems, parts &amp; sub systems development</td>
<td>• Telecom Engineering</td>
<td>• Infrastructure &amp; civil works</td>
</tr>
<tr>
<td>• Electronics</td>
<td>• Modelization, simulation &amp; testing</td>
<td>• Media Engineering</td>
<td>• Process &amp; facilities</td>
</tr>
<tr>
<td>• Software engineering</td>
<td>• Process, industrialization &amp; tooling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Safety</td>
<td>• Material development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Security</td>
<td>• Mechanical specialties</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Connectivity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Energy &amp; Environment engineering</th>
<th>Lifescience &amp; healthcare engineering</th>
<th>Finance banking &amp; Insurance expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Energy &amp; environment management</td>
<td>• Healthcare specialties</td>
<td>• Specific finance bank &amp; insurance specialties</td>
</tr>
<tr>
<td>• Nuclear Engineering</td>
<td>• Lifescience specialties</td>
<td></td>
</tr>
<tr>
<td>• Fossil Fuel, oil &amp; gas Engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Renewable Energy Engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Electric network</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Competence lines
Disciplinary fields
Proposed Knowledge Graph

**Label** - part of a group:
- *Competence Keywords*

**Relationship type:**
- *Has*
- *Is a*

**Scale-free network** - a scale-free network is produced when there are power-law distributions and a hub-and-spoke architecture is preserved regardless of scale, such as in the World Wide Web.

**Flavors of Graphs:**
- *Disconnected*
- *Unweighted*
- *Tree*
- *Sparse*
- *Bipartite – function and specialty*
Proposed approach of search engine

In the search engine, we take two kinds of files as input – CV and clients inquiries. We make a pre-processing to remove the ambiguous (for example, “JS” is “Java Script”). Then we use NER tool based on DistilBERT-base-multilingual-case to extract the competence keywords and TF-IDF to calculate the score of the competence keywords. We also use Weight Average Method (WAM) to calculate a global score. The scores are stored in Redis with index and inverted index. We select a list of candidates with high score as well as the related competence keywords from the Knowledge Graph.
Term Frequency–Inverse Document Frequency (TF-IDF)

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.

\[ w_{i,j} = tf_{i,j} \times idf_i \]

\[ idf_i = \log \left( \frac{1+N}{1+df_i} \right) \]

- \( w_{i,j} \) = score of term i in document j
- \( tf_{i,j} \) = number of occurrences of term i in document j
- \( df_i \) = number of documents containing the term i
- \( N \) = total number of documents
Weight Average Method

Propose the weighting to calculate an overall score for the four sections of the CV - experience, core skill, skill keyword and activity keyword.

Defuzzification methods:
• Center of gravity
• Centroid method
• Center of sum
• Max-membership principal
• Weight Average Method

\[ Z = \frac{\sum \mu(z) \cdot z}{\sum \mu(z)} \]
Scientific approach for the search engine

Altran Official

Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT
Experimentation results – search candidates by « Java »

It is obvious to see that Linx does not pay attention to the experiences of candidates. All the three candidates from BERT have experiences of Java and this skill is also highlighted on the section of core skill and skill keyword. Especially, the first candidate from BERT has the experience of “Eclipse” which is related to “Java” in the knowledge graph.

<table>
<thead>
<tr>
<th>ID candidate</th>
<th>Experience</th>
<th>Core Skill</th>
<th>Skill Keyword</th>
<th>ID candidate</th>
<th>Experience</th>
<th>Core Skill</th>
<th>Skill Keyword</th>
</tr>
</thead>
</table>
| 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, Go 1.6, Java, TDD | 346575  
Developer at AIRBUS - France: Creation of a Platform for Icing studies Development at AIRBUS - France: Prer evolution 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 |
| 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, SSD, Html5 |  |
| 15868        | Tester at HP ENTERPRISE SERVICES COMMUNICATION & MEDIA SOLUTIONS - France  
System Tests and Functional Tests on a virtualized system of the 4G network 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 | Functional testing and validation, Test and technical validation, Collaboration and networking | Collabnet Svn, Teamforge Svn, Microsoft Office, HP ALM, Collaborative Tools | 67711  
Developer at ACS - France: OA: The system quantifies the fatigue at work for an employee during his working hours,  
Developer at ACS - France: 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. |  | Application WEB, Core network mobile circuits, Product design and development | HTML, JavaScript, AngularJS, Angular, Java |
Perspectives

• An advanced word embedding tool is required to replace TF-IDF Vectorizer.

• A richer KG is needed with the properties of each node and weighted relation for the dynamic management.

• A recommendation system with KG embedding method is needed for the matching between the profile of candidate and the description of mission.
Thank you for your attention!

Yan WANG, Yacine ALLOUACHE, Christian JOUBERT
Capgemini Engineering, Direction of Research and Innovation (DRI)
yan.wang2@altran.com, yacine.allouache@altran.com, christian.joubert@altran.com