Beyond generic skills: creating capability-centric, company-specific knowledge graphs from job descriptions

Kaan Karakeben[1], Henri Egle Sorotos[2] and Ahmad Assaf[3]

[1]Beamery, London, United Kingdom
[2]Beamery, London, United Kingdom
[3]Beamery, London, United Kingdom

1. Motivation

The Human Resources domain is a data-rich function in any organisation. Semantic knowledge graphs have previously been discussed as a key enabler to bring clarity to the data, and to enable better decision making. Specifically, a previously proven solution to extract a limited number of basic entities from job descriptions has been established [1]. The previous method linked to a standardised taxonomy of skills, and the result was used to generate Resource Description Framework (RDF) graphs. Additional challenge of this was that these approaches were limited to skills only, and do not take into account the full context of a company [1] [2]. The resulting graph is often generic, incomplete, and not useful for human decision making.

In reality, a job description consists of a set of capabilities describing an ideal candidate for the post. Capabilities act as an intermediate layer between a job description's responsibilities and the skills they required to fulfill them. They encompass the common tasks within responsibilities and help identify relevant skills. By doing so, they incorporate the company context into the skills derived from these responsibilities.

This paper introduces a novel approach to address the limitations of existing methods by extracting company-specific and bespoke capabilities. The result is a rich graph of both generic and bespoke entities, which can be used to generate a more complete and useful RDF graph. Outcomes are grounded in the context of the company.

2. Solution

The solution is a multi-step process, which extracts entities from job descriptions iteratively. Entities are extracted in a hierarchical manner, starting with the most conceptually aggregated entities, and moving to the most granular. Each stage involves extracting more granular entities from entities extracted in the previous stage.

For example, a set of capabilities and responsibilities are extracted as part of the initial stages, and these are then used to extract skills and others in the later stages. This is depicted in figure 1.

As entities become more granular, they become less specific to the company. For instance, the capabilities and responsibilities are highly company specific, but also conceptually aggregated and abstract. They may contain references to specific technologies used, processes followed, or attributes of the company culture. These company specific entities have a high cardinality. It is extremely unlikely that two companies will have a large overlap in these entities. As entities become more granular, they become less specific to the company. Skills are the final stage of extraction, and are a common currency across companies. There will be a high degree of overlap in the skills extracted from different

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA
kaan.karakeben@beamery.com (K. Karakeben); henri.egle-sorotos@beamery.com (H. E. Sorotos); ahmad@beamery.com (A. Assaf)
www.beamery.com (K. Karakeben); www.beamery.com (H. E. Sorotos); www.beamery.com (A. Assaf)

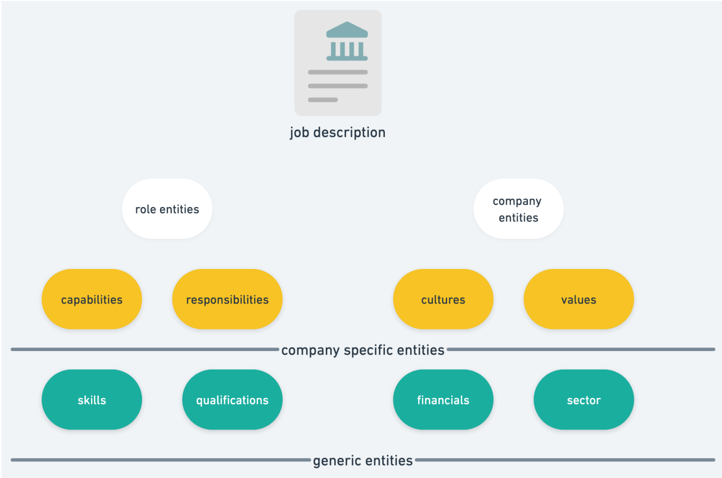

Figure 1: Domain Model for extraction

companies. These generic skills are then linked to a proprietary taxonomy of skills contained in the Beamery Knowledge Graph.

As well as extracting entities specific to the role, an additional benefit of the method is extracting entities describing the company itself. These include, but are not limited to, company values, culture and technologies. Additionally, attributes of the legal entity such as company age, size, sector and revenue can also be found within the job descriptions. This represents a new but adjacent domain of entities that can be extracted. These raw entities are then linked to a standardised taxonomy of company-specific entities contained within Beamery's proprietary taxonomy. Taxonomies are all stored using agreed semantic web standards [4], enabling simple linking and downstream use.

The result is multiple interlinked adjacent domains of entities.

The extraction process is facilitated by a bespoke Large Language Model (LLM) system that has been built to specifically extract entities from job descriptions [1].

The system is prompted to process a custom extraction with clear assumptions and expectations from its concepts unlike generic LLM based systems, and also facilitates automatic reconciliation with existing canonical entities within the Beamery Knowledge Graph [4] [3].

The resulting graph, which matches the precision of human annotators, enriches downstream tasks with additional entities and company-specific context. This contextual alignment has significantly accelerated user adoption of AI features, as the incorporation of company-specific details into the explanations has made the user experience more intuitive and relevant.

3. Conclusions

The major shortcoming of the previous method is that it is limited to skills. Whilst this meant AI systems could understand these entities, it also meant that the resulting graph was generic and not useful for decision making. Humans consumers found little value in the graph, as it did not reflect the reality of the company. Downstream tasks such as job matching, skills gap analysis, and workforce planning can benefit from the richer graph. This is already deployed and in use at Beamery.

Whilst job descriptions are a rich source of data, they are not the only source. Future developments are likely to focus on extracting similar entities from other artefacts, such as resumes, competency frameworks, and company reports.

References

[1] Kaan Karakeben, Henri Egle Sorotos, Alisa Milchevskaya, and Ahmad Assaf, "Rapid Graph Generation from Job Descriptions: Combining Taxonomies and LLMs," 2024.

[2] M. Zhang, K. N. Jensen, S. D. Sonniks, and B. Plank, "SkillSpan: Hard and Soft Skill Extraction from English Job Postings," *arXiv preprint*, 2022. arXiv:2204.12811.

[3] L. Dodds and I. Davis, *Linked Data Patterns: A pattern catalogue for modelling, publishing, and consuming Linked Data*, 2022. Retrieved from https://patterns.dataincubator.org/book/.

[4] W3C, "SKOS Simple Knowledge Organization System Reference," 2009. Retrieved from https://www.w3.org/TR/skos-reference/.