Leveraging Textual Information for Knowledge Graph-oriented Machine Learning: A Case Study in the Construction Industry

Shahinmoghadam M. [1], Motamedi A. [1], Soltani M.M. [2]
[1] École de technologie supérieure, Canada, [2] Toronto, Canada
Mehrzad.Shahinmoghadam.1@ens.etsmtl.ca, Ali.Motamedi@etsmtl.ca, mo_solta@encs.concordia.ca

Abstract. The proven power of Knowledge Graphs (KGs) to effectively represent lexical and semantic information about numerous and heterogeneous entities and their interconnectedness has led to the growing recognition of their potential in engineering disciplines. Meanwhile, a greater focus has been placed on graph embedding techniques to derive dense vector representations of KGs. Such representations could enable the use of conventional machine learning techniques over the content of KGs. However, in the context of building engineering, the quality of the graph embeddings could be problematic, mainly due to the relatively small size of the KGs that are created for individual buildings. This paper aims to investigate the effectiveness of applying KG embedding methods when the elements of the building are described narrowly within the KG. The results of our experiments confirm that proper use of data transformation techniques can significantly improve the quality of the feature representation for downstream tasks.

1. Introduction

Due to the ubiquitous presence of graph (network) structures in many real-world phenomena, graph-oriented analytics has been the subject of interest in many fields of research (Goyal and Ferrara, 2018). Recently, both academia and industry practitioners have shown an increased interest in the notion of Knowledge Graphs (KGs). In general terms, a KG can be defined as a structured representation of facts about real-world or abstract subjects (Ji et al., 2022). Each KG is composed of nodes and edges (entities and relationships) which jointly represent facts (semantic descriptions) about different subjects in an explicit and machine-readable format.

Successful applications of Knowledge Graphs (KGs) for purposes including but not limited to context-aware information retrieval and question-answering over knowledge bases, has attracted considerable attention in various engineering disciplines. The construction and building engineering sector have been no exception to this trend. The added value of KGs and their potential for open exchange and management of the built environment data has been most recently highlighted in (Pauwels et al., 2022).

Despite their vast expressive power to represent knowledge in a structured way, manipulation of KGs is challenging for purposes such as statistical learning (Wang et al., 2017). To address these challenges, a fast-growing volume of research has been dedicated to represent the entities of the KG in the form of numerical vectors. In the context of graph analytics, this approach is known as graph embedding. Representation of the KG components in low-dimensional vector spaces will significantly facilitate applicability of the machine learning methods over the content of the KG. This way, further knowledge can be extracted from KGs when symbolic inference mechanisms are incapable of or inefficient in doing so.

The effectiveness of the existing graph embedding techniques has been under investigation in various domains (e.g., bioinformatics, social network analysis) and the results have been encouraging (Goyal and Ferrara, 2018). Moreover, researchers have proposed novel embedding techniques that are customized to specifically cope with the content of KGs. However, the majority of the existing embedding techniques have been tested with large-scale graphs that usually contain billions of facts (Rossi et al., 2021), while the average size of a KG that is

https://doi.org/10.7146/aul.455.c216

constructed for an individual building (or small group of buildings) will be significantly smaller in scale. Moreover, the structural pattern of a building KG may be quite different from open-domain KGs such as Wikidata. Hence, the applicability of the KG embedding techniques in the context of building engineering remains questionable.

This paper aims to address the above-mentioned issue by examining the quality of the results of applying existing KG embedding algorithms over small-sized building KGs. To meet this objective, a case study approach was adopted. First, a working dataset was created whose content was extracted from existing reference models, which had originally been developed for building engineering intentions. Subsequently, a set of experiments were carried out with the purpose of deriving the embeddings of the building graph to be used as the input to a node classification model. The results of our experiments showed that despite the possibility of encountering challenges within the downstream machine learning phase (e.g., reduced applicability of linear models), careful curation of the embeddings (e.g., kernel approximation) can significantly improve the effectiveness of the derived feature vector for supervised/unsupervised learning. The remainder of the paper describes the required background, methods used, and experiments conducted, followed by the discussion and conclusion.

2. Background

2.1 KG definition and characteristics

Following the recent successful adoption of KGs in various domains of practice, the volume of research publications on KGs has been increasing (Hogan et al., 2021). Despite the previous research efforts made to provide a widely-accepted formal definition of KGs, no such definition currently exists within the literature (Ji et al., 2022). Yet, in a general sense, a KG can be viewed as a graph composed of nodes and edges, which together provide formal descriptions of the entities of interest and their interconnectedness (Ehrlinger and Wöß, 2016; Hogan et al., 2021). In some of the reference publications, the incorporation of the KG content into formal ontologies and the use of the semantic reasoners has been essential to the definition of KGs (Ehrlinger and Wöß, 2016). Thanks to the reasoning power of formal ontologies, numerous implicit relationships can be automatically inferred and added to the original KG as new factual statements. Regardless of the way a KG is created, one of its outstanding potentials is its ability to preserve the "context" in data representation. The explicit representation of contextual correspondences within a KG provides the possibility of interpreting the data from different perspectives (Hogan et al., 2021). This is of great value to the development of various types of context-aware applications (information/knowledge retrieval, recommender systems, industrial chatbots, etc.)

2.2 KGs in the built environment

Most recently, Pauwels et al. (Pauwels et al., 2022) elaborated on the notion of KGs for the built environment. In their work, researchers put a significant weight on the principles of "linked data" and "semantic web" technologies, e.g., OWL (Web Ontology Language (McGuinness and Van Harmelen, 2004)) and RDF (Resource Description Framework (Lassila and Swick, 1998)). Thanks to the availability of the mature ontologies that currently exist for the built environment, KGs can be constructed practically to represent facts about the design, construction, and operation of the real-world built entities. The nodes of such KGs can refer to both real-world entities (e.g., a physical space) or virtual ones (e.g., average floor temperature

sensor (Balaji et al., 2018)), while the edges of the KG make reference to the semantic relationships that exist between the entities of the building. When incorporated into formal ontologies such as ifcOWL (buildingSMART, 2022), each pair of nodes linked via an edge (known as a triple) expresses a "fact" about the building, in an explicit and formal (machine-readable) manner. However, the inclusion of detailed geometric descriptions as well as sensory observations in KGs is associated with considerable complexities and inefficiencies (Pauwels et al., 2022). Hence, from a practical perspective, the content of the KGs that are constructed for the built environments will mainly consist of lexical and semantic descriptions. Nevertheless, a large amount of valuable (semantic) information can be still represented within KGs that will be of great value to knowledge extraction purposes.

2.3 KG representation learning

The key motivation behind our current study is that with the help of the "semantic embedding" techniques, the contextualized information contained in the built environment KGs can be exploited for knowledge extraction using machine learning methods. Semantic embedding can be defined as encoding the semantics of the data (KG content in our case) into numeric vector representations (Bengio et al., 2013). In other words, KG embedding techniques can be used to generate a feature vector by which the textual information contained in KGs can be represented in a numeric form. Recently, there has been a growing interest in developing semantic embedding techniques that can effectively cope with the particular characteristics of KGs (i.e., dealing with lexical and semantic information). One of the relatively recent algorithms in this regard is RDF2Vec (Ristoski et al., 2019), which was originally proposed to learn vector representation of RDF-formatted KGs. However, the effectiveness of the existing techniques has been mostly tested over large-scale open-domain KGs (e.g., Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014)), and rarely investigated in construction and building engineering research.

2.4 Research gap and objectives

Given the relatively significant small size of the KGs that are created for individual built environments, compared to the size of the open-domain benchmark KGs, the applicability of the existing KG-embedding techniques for the built environment needs to be carefully investigated. This study set out to contribute to filling this gap by focusing on the RDF-formatted KGs that are built for individual buildings. Our main objective is to investigate the applicability of the RDF2Vec method by examining the usability of the generated embeddings in a node classification scenario.

3. Methods and materials

3.1 Data preparation

To create a working dataset for the purpose of this study, we used two of the reference building models that are publicly available at ("BrickSchema.org", 2022). These reference models, serialized in RDF format, are primarily intended for building engineering research purposes. They are representative examples of the use of the Brick schema (ontology) to deliver rich semantic descriptions of a building's physical, logical and virtual assets and their relationships ("BrickSchema.org", n.d.). Among the five reference models that were available at the time of our study (early January 2022), we selected the "Gates Hillman Center" (GHC) and "Engineering Building Unit 3B" (EBU) models, which contained approximately 35,000 and

8,000 relationships, respectively. The percentage of mapping the building's actual data points to the entities of the used schema (Brick ontology) for the GHC and EBU models were 99% and 96%, respectively. The rationale behind the selection of these two models was to reinforce the validity of the results by performing our experiments using two datasets that were different in size. A summary of the description of the models is provided in Table 1. More details on the characteristics of the actual buildings and their associated semantic model can be found in (Balaji et al., 2018).

Table 1: Summary of the description of the data.

General description	Description of the semantic models (KGs)			Description of the created datasets (Target class statistics)			
	Graph nodes	Total relationships	Unique relationships	Train/Test samples	Point/ subclass	Location/ subclass	Equipment/ subclass
Gates Hillman Center (GHC); 217k (ft^2) floor space, Built 2009	≈ 9.6k	≈ 35.7k	9	1452 364	934 Sensor	475 Room	447 VAV (variable air volume)
Engineering Building Unit 3B (EBU); 150k (ft^2) floor space, Built 2004	≈ 6.1k	≈ 8.4k	4	576 145	238 Sensor	246 Room	237 Damper

Several SPARQL queries were conducted, over the two models mentioned above, to create two distinct datasets for the entities of interest, i.e., the entities for which the embeddings were calculated. Three distinct types, namely "Point", "Location", and "Equipment", were considered as the target classes for the development of the datasets. It should be noted that for the "type" relationships, the original models only contained triples that indicated the subclasses of the three mentioned target classes. For example, while the model contained type relationships indicating the "Room" class, the fact that a room is a subclass of the "Location" class was absent from the model. Finally, each created dataset was split into a training set (80% of samples) and test set (20% of samples). The summary of the statistics for each created dataset can be seen in Table 1.

3.2 Approach

A summary of our approach towards the experimental evaluations conducted in this study is shown in Figure 1. Description of the three main modules depicted in the figure is given as follows:

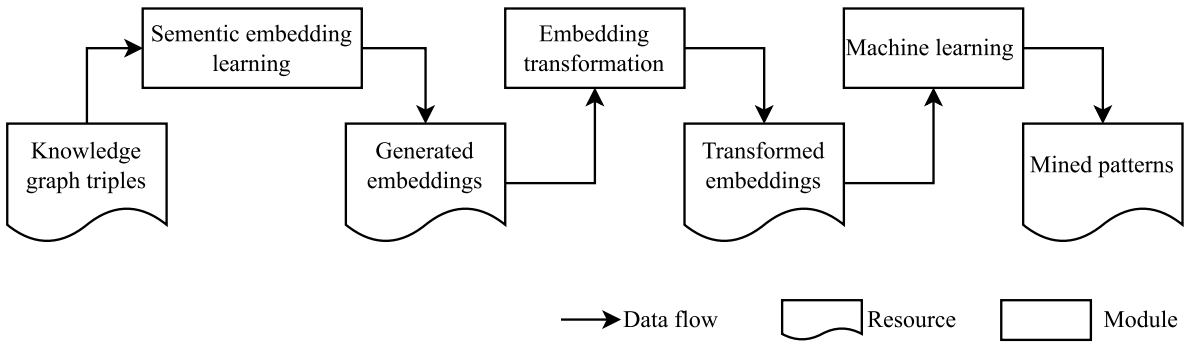

Figure 1: Proposed approach.

Semantic embedding learning. As mentioned in the previous section, we used the RDF2Vec algorithm to generate the embeddings for the instances of points, locations, and equipment in the created datasets. A Python implementation of RDF2vec (Vandewiele et al., 2020) was used in this study to calculate the KG embeddings. For all experiments, the size of the embeddings was set to 100. In other words, at each round of embedding calculation, a feature vector with 100 dimensions was learned to represent the entities of interest in a continuous vector space.

Embedding transformation. In the proposed approach, the embedding transformation module is responsible for maintaining the applicability of the generated embeddings for downstream tasks such as node classification as intended in the current study, as well as other tasks (e.g., exploratory data analysis, link prediction, etc.). Both linear and non-linear data transformation methods were considered for our embedding transformation module. In particular, we considered the following algorithms:

Principal Component Analysis (PCA). As one of the most popular dimensionality reduction techniques, PCA was considered for the removal of those dimensions from the generated embeddings that express insignificant rates of explained variance among the data points. Using PCA, non-representative dimensions can be removed from the feature vector (noise reduction), which in turn can improve the overall computational performance for downstream machine learning. Moreover, 2D/3D visualizations of the embeddings based on the identified principal components can be used to intuitively observe the discriminative power of the learned vector representation.

Kernel Principal Component Analysis (Kernel PCA). Despite its recognized strengths, there are many probable occasions in which linear PCA fails to find the linear separation between the dissimilar instances of the data. To tackle this issue, we considered the use of kernel approximation methods to find a projection of the embeddings (in higher dimensions) that allows for them to be linearly separated.

t-SNE (t-distributed Stochastic Neighbor Embedding). T-SNE is highly effective for the 2D/3D visualization of high-dimensional data (Van der Maaten and Hinton, 2008). Given the relatively large size of the embeddings generated for each dataset (100 dimensions), we first used the t-SNE algorithm to intuitively observe the distance at which the dissimilar embeddings (node types) were clustered apart from each other.

Node classification. The ultimate goal of our node classification module is to assign a distinct "type label" to an unseen building node entity. In particular, the objective is to construct a model that takes the transformed (pre-processed) embeddings of a node that is a joint product of the two previous modules as input and gives a label of the class to which the node belongs to (object type) as output. With reference to the experimental setup that was used in this study for dataset creation, each node can only belong to one of the "Point", "Location", or "Equipment" classes.

For the sake of model development, we tested the predictive performance of three different algorithms each from a different group of supervised learning methods. In particular, we used "Logistic Regression", "Random Forest", and "Multi-layer Perceptron" algorithms, which belong to linear, ensemble, and neural network machine learning models, respectively. For the sake of the evaluation metrics, we computed "accuracy score" and "F1-score" (harmonic mean of precision and recall) to compare the prediction performance of each trained model.

4. Results and discussion

4.1 Experiments

Before we address the experimental results, it should be noted that the codes and created datasets will be made available by request, for the benefit of the community and to ensure the reproducibility of the reported results. Moreover, it should be mentioned that the main purpose of this case study has been to highlight the potential benefits of KG embeddings for machine learning purposes in an industrial setting. Hence, an in-depth comparison of the existing techniques to derive and transform the embeddings, as well as hyper-parameter tuning for the methods used in the present work will be left to future research.

In the first step of our experimental evaluations, we calculated the embeddings for the two datasets that were created for the purpose of this study. Figure 2 shows the results of the use of t-SNE to deliver 2D visualizations of the generated embeddings for GHC and EBU datasets, respectively. A closer look at Figure 2 reveals two key observations: First, the embeddings derived for none of the datasets were clustered quite far apart, thereby restraining the linear separation of the dissimilar classes. Second, while the distribution of the GHC embeddings follows a perceptible (circular) pattern, the embeddings generated for the EBU dataset seem to be more arbitrary.

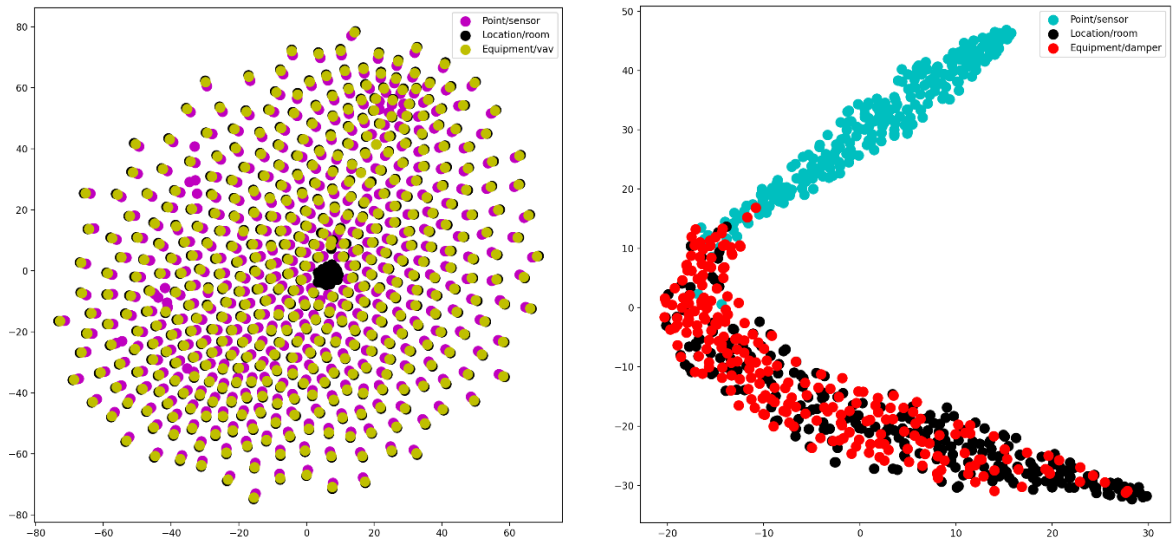

Figure 2: t-SNE visualization of embeddings for GHC (left) and EBU (right) datasets.

To find a plausible explanation for the dissimilar behaviour of the two datasets, we looked at the underlying structure of the KGs upon which the embeddings were generated. As one can see from Table 1, while 9 unique relationships can be found in the GHC model, only 4 unique relationships were used to describe the entities of the EBU semantic model. Moreover, the ratio of total relationships to nodes, for the GHC model (≈ 35.7/9.6), is almost three times higher than that of the EBU model (≈ 8.4/6.1). Based on these comparisons it can be argued that the granularity of the semantic descriptions within the GHC model has been higher, comparatively. To further reinforce our argument, we conducted an exploratory analysis by running multiple queries (SPARQL queries) over the reference models (KGs). The results of our analysis confirmed that the structure of the description of the entities in the used KGs is comparatively different from each other, in terms of both syntactic and semantic characteristics. For example, while up to 6 unique relationships were used to describe the room entities inside the GHC model, only 2 relationships were used within the EBU model for the same purpose. Hence, it

can be concluded that different "modelling patterns", i.e., using different terminology and structures to describe the entities of a building within a KG can lead to varying behaviours of the derived embeddings. However, answering the question of which modelling pattern leads to better data representation requires profound research considerations and is beyond the scope of the present study.

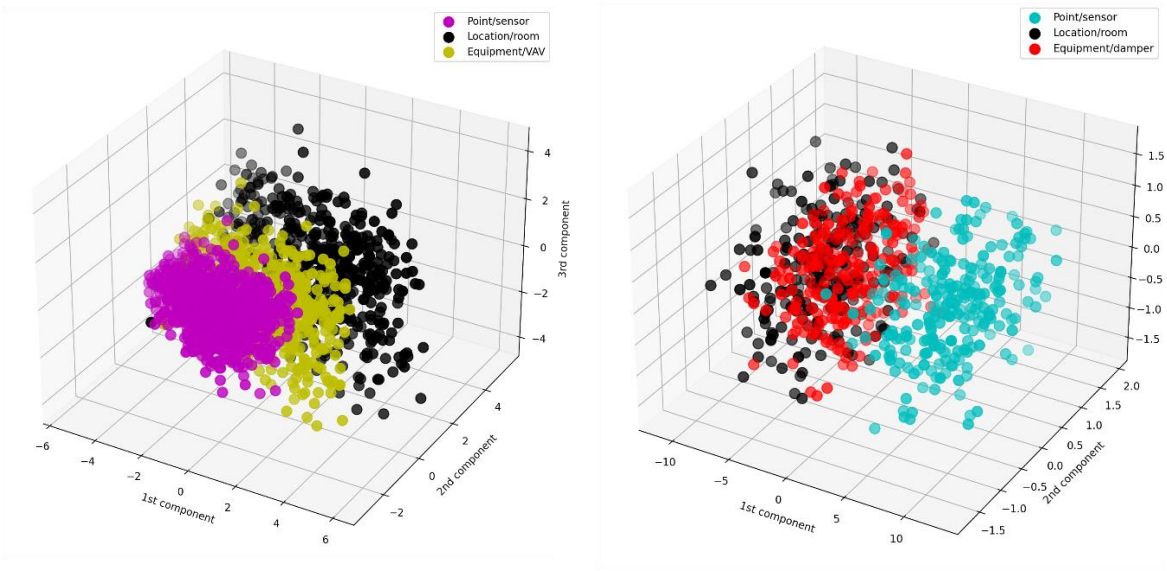

Figure 3: PCA-based transformation of embeddings for GHC (left) and EBU (right) datasets.

As mentioned in section 3.2, we considered the use of both linear and non-linear projection methods to improve the quality of the feature representation. The plots depicted in Figure 3 and Figure 4 show the results of the use of linear and kernel PCA, respectively, to deliver 3D visualizations of the transformed embeddings. As one can see from the comparison of the plots in Figure 3 and Figure 4, the used methods show competitive discriminative power. To quantitatively compare the quality of the feature representations, we calculated the variance of the projected samples. This step was taken to identify feature representation with lower variance as it points to a denser representation, which will be more favourable for statistical learning purposes. The results confirmed that for both datasets, kernel approximation outperformed linear PCA, in terms of delivering dense representations (see Table 2).

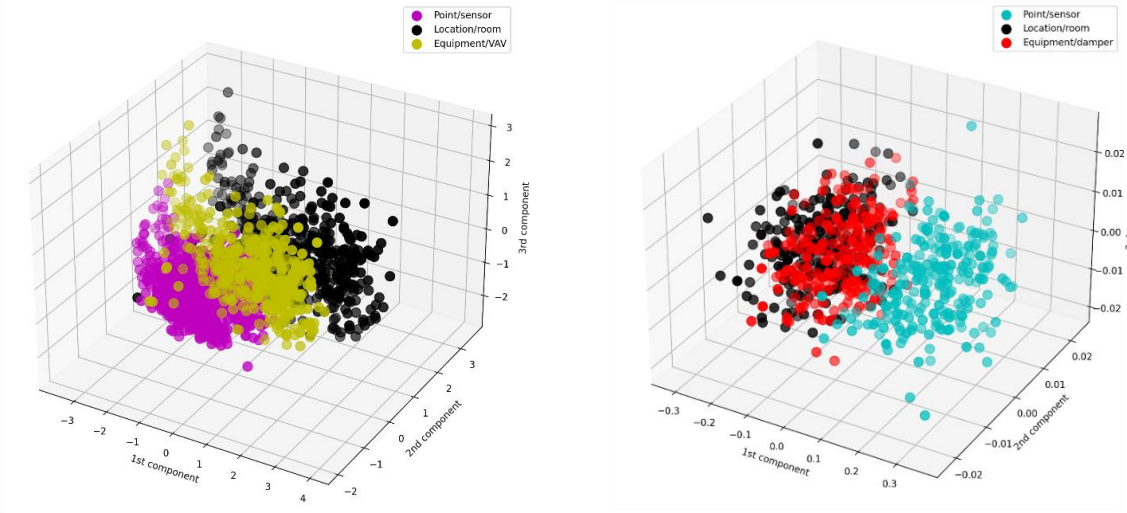

Figure 4: Kernel approximation of embeddings for GHC (left) and EBU (right) datasets.

Table 2: Variance of the transformed embeddings calculated for 1st to 3rd principal components.

	Linear PCA	**Kernel PCA**
GHC dataset	2.96	1.21
EBU dataset	9.9	0.01

Subsequent to obtaining encouraging results with the use of kernel PCA to find the projection of the generated embeddings in higher dimensions, the next step was to perform a dimensionality reduction step. To identify the effective number of principal components to be used for the representation of the training data, we used a base classifier (i.e., a Random Forest model with 100 decision trees) and observed the quality of the predictions as we changed the number of the components. As a result of taking this step, the size of the feature vector was set to 80 and 100 for the GHC and EBU samples, respectively. Hence, prior to the training and testing of our node classification models, we reduced the dimension of the feature vectors (embeddings) to decrease noise and improve the computational performance. After training three distinct classification models for each dataset, the test sets were introduced to the developed models to evaluate the quality of their predictions. The prediction performance scores of the trained models are presented in Table 3.

Table 3: Summary of the node classification results.

	Tested with GHC dataset			**Tested with EBU dataset**		
	Random Forest	**Logistic Regression**	**ML-Perceptron**	**Random Forest**	**Logistic Regression**	**ML-Perceptron**
Accuracy	0.91	0.96	0.95	0.76	0.76	0.79
F1-score	0.91	0.96	0.95	0.76	0.75	0.78

It is apparent from Table 3 that the accuracy of the predictions made by the models that were trained with the GHC data is higher than that of those made by the models that were trained with the EBU dataset. The different size of the data used to train the two sets of the models can be mentioned as an immediate explanation for this observation. However, with reference to our earlier discussion of the effect of the structural differences in KG creation (e.g., granularity of the semantic descriptions), it is strongly possible that the lower performance of the models that were trained with EBU data, stems from the quality of the content of the KG.

4.2 Implications to research and practice

The results of testing our trained models confirmed that relatively accurate predictions of object types could be realized based purely on the embeddings of the lexical and semantic information that exist for a building entity inside the KG. Given the small size of the used KGs, in addition to the narrow semantic descriptions that existed for the building entities inside the KGs (particularly for the case of the EBU model), it can be concluded that even minimalistic semantic descriptions inside the individual building KGs can effectively contribute to the identification of latent semantic patterns. These findings can be of important value to various lines of research in the built environment. In particular, the research on the notion of Building Information Modelling (BIM), which has been traditionally abundant with rule-based approaches, can be advanced with the help of KG-oriented machine learning techniques for context-aware machine learning. In this respect, semantic enrichment of the building information models is one of the areas with the most potential in which KG embeddings can be

used to facilitate machine learning from graph-structured linked sources of building data. Among other worthwhile applications, context-aware information retrieval and Question-Answering systems for the built environment can benefit from KGs and their embeddings. In fact, the learned embeddings could be used for the unsupervised clustering of the content of the building KGs. Then, by assigning relevant semantic tags to the identified clusters, the required information can be searched through contextually relevant information that is contained inside the most relevant clusters. Moreover, given the importance of query relaxation/approximation for effective discovery of meaningful information from semantic building models (Bennani et al., 2021), KG embeddings can be of value in this regard, as previous research has shown encouraging results in other domains (Mai et al., 2020; Wang et al., 2018).

Finally, the key implications and values of our findings to research and practice can be summarized as follows: To the best of our knowledge, this work is among the very few existing papers that investigate the effectiveness of KG embedding techniques in the context of building engineering, while these techniques can have considerable implications in the realm of building analytics. In fact, while KGs can provide valuable background knowledge about the building, the corresponding embeddings provide ready-to-use numeric representations of that knowledge in continuous vector spaces. This allows the background knowledge about the building to be incorporated for downstream machine-learning tasks. Most importantly, with the absence of the building's geometry and operational time-series data in the KG (see section 2), entity embeddings can be used to generate vector representations of the lexical and semantic information that is contained in the KG, thereby facilitating "context-aware building analytics".

5. Conclusion

This study set out to investigate the usefulness of applying KG embedding techniques to individual building KGs. Our results provided quantitative evidence that semantic representation learning techniques in combination with careful pre-processing of the learned embeddings can enable the use of machine learning techniques to find latent semantic patterns from the lexical content of the building KGs. We also found that the modelling pattern used to create a building KG, i.e., word synthesis and choice of the semantic relationships, is an important determinant of the quality of the generated embeddings (feature vector). Hence, the identification of best modelling practices for KG creation for individual buildings is an important direction for future research. Another important direction for future research would be to investigate the effectiveness of KG embedding methods for applications such as context-aware information retrieval and Question-Answering from building KGs.

References

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y. and Bergés, M. (2018), "Brick: Metadata schema for portable smart building applications", Applied Energy, Vol. 226, pp. 1273–1292.

Bengio, Y., Courville, A. and Vincent, P. (2013), "Representation Learning: A Review and New Perspectives", IEEE Transactions on Pattern Analysis and Machine Intelligence, presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35 No. 8, pp. 1798–1828.

Bennani, I.L., Prakash, A.K., Zafiris, M., Paul, L., Roa, C.D., Raftery, P., Pritoni, M. and Fierro, G. (2021), "Query relaxation for portable brick-based applications", Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Association for Computing Machinery, New York, NY, USA, pp. 150–159.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. (2008), "Freebase: a collaboratively created graph database for structuring human knowledge", Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, New York, NY, USA, pp. 1247–1250.

"BrickSchema.org". (2022), available at: https://brickschema.org/ (accessed 23 February 2022).

buildingSMART. (2022), "ifcOWL", BuildingSMART Technical, available at: https://technical.buildingsmart.org/standards/ifc/ifc-formats/ifcowl/ (accessed 23 February 2022).

Ehrlinger, L. and Wöß, W. (2016), "Towards a definition of knowledge graphs", SEMANTiCS (Posters, Demos, SuCCESS), Citeseer, Vol. 48 No. 1–4, p. 2.

Vandewiele, G., Steenwinckel, B., Agozzino, T., Weyns, M., Bonte, P., Ongenae, F., & De Turck, F. (2020), "pyRDF2Vec: Python Implementation and Extension of RDF2Vec", IDLab, https://github.com/IBCNServices/pyRDF2Vec. Goyal, P. and Ferrara, E. (2018), "Graph embedding techniques, applications, and performance: A survey", Knowledge-Based Systems, Vol. 151, pp. 78–94.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S. and Ngomo, A.C.N.,. (2021), "Knowledge Graphs", ACM Computing Surveys, Vol. 54 No. 4, p. 71:1-71:37.

Ji, S., Pan, S., Cambria, E., Marttinen, P. and Yu, P.S. (2022), "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications", IEEE Transactions on Neural Networks and Learning Systems, presented at the IEEE Transactions on Neural Networks and Learning Systems, Vol. 33 No. 2, pp. 494–514.

Lassila, O., & Swick, R. R. (1998), "Resource description framework (RDF) model and syntax specification".

Mai, G., Yan, B., Janowicz, K. and Zhu, R. (2020), "Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model", in Kyriakidis, P., Hadjimitsis, D., Skarlatos, D. and Mansourian, A. (Eds.), Geospatial Technologies for Local and Regional Development, Springer International Publishing, Cham, pp. 21–39.

McGuinness, D. L., & Van Harmelen, F. (2004). "OWL web ontology language overview", W3C recommendation, 10(10), 2004.

Pauwels, P., Costin, A. and Rasmussen, M.H. (2022), "Knowledge Graphs and Linked Data for the Built Environment", in Bolpagni, M., Gavina, R. and Ribeiro, D. (Eds.), Industry 4.0 for the Built Environment: Methodologies, Technologies and Skills, Springer International Publishing, Cham, pp. 157–183.

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R. and Paulheim, H. (2019), "RDF2Vec: RDF graph embeddings and their applications", Semantic Web, IOS Press, Vol. 10 No. 4, pp. 721–752.

Rossi, A., Barbosa, D., Firmani, D., Matinata, A. and Merialdo, P. (2021), "Knowledge Graph Embedding for Link Prediction: A Comparative Analysis", ACM Transactions on Knowledge Discovery from Data, Vol. 15 No. 2, p. 14:1-14:49.

Van der Maaten, L. and Hinton, G. (2008), "Visualizing data using t-SNE", Journal of Machine Learning Research, Vol. 9 No. 11.

Vrandečić, D. and Krötzsch, M. (2014), "Wikidata: a free collaborative knowledgebase", Communications of the ACM, ACM New York, NY, USA, Vol. 57 No. 10, pp. 78–85.

Wang, M., Wang, R., Liu, J., Chen, Y., Zhang, L. and Qi, G. (2018), "Towards empty answers in SPARQL: approximating querying with RDF embedding", International Semantic Web Conference, Springer, pp. 513–529.

Wang, Q., Mao, Z., Wang, B. and Guo, L. (2017), "Knowledge Graph Embedding: A Survey of Approaches and Applications", IEEE Transactions on Knowledge and Data Engineering, presented at the IEEE Transactions on Knowledge and Data Engineering, Vol. 29 No. 12, pp. 2724–2743.