Knowledge Graph Embedding in E-Government

Friday Orji*
Department of Computer Science
Rivers State University, Port Harcourt, Nigeria
friday.orji@ust.edu.ng

Okoni Bennett
Department of Computer Science,
Rivers State University, Port Harcourt, Nigeria
bennett.okoni@ust.edu.ng
DOI: 10.56201/ijcsmt.v11.no2.2025.pg97.105

Abstract

The increasing adoption of Knowledge Graphs (KGs) ***in*** *e-Government presents new opportunities for improving knowledge representation, inference, and decision-making in digital governance. However, real-world knowledge graphs often suffer from incompleteness, limiting their effectiveness in tasks such as policy analysis, citizen engagement, and service integration. To address this challenge, we explore Knowledge Graph Embedding (KGE), a Machine Learning (ML) technique, to generate meaningful vector representations of entities and relationships within an RDF/OWL knowledge graph developed using Protégé. Specifically, we apply TransE, a simple yet effective translational embedding model, to encode structured e-Government data and predict missing links. Our approach involves preprocessing the knowledge graph using RDFLib, extracting entity-relation triples, and leveraging PyKEEN for embedding computation and evaluation. The dataset is divided into 80% training, 10% validation, and 10% test sets, and model performance is assessed using standard metrics, including Mean Rank (MR), Mean Reciprocal Rank (MRR), and Hits@k (k = {1, 3, 10})***.*** *Experimental results reveal structural patterns within the e-Government knowledge graph, highlighting key entity clusters, relationship distributions, and link prediction trends. Furthermore, our analysis indicates that tail entity prediction tends to be more reliable than head entity prediction, suggesting areas for further optimization. Although TransE demonstrates promising results, challenges such as high variance in ranking metrics, dataset sparsity, and potential biases in entity distribution remain. Future research directions include hyperparameter tuning, integration of external ontologies, and the incorporation of neuro-symbolic AI techniques to enhance semantic reasoning. This study contributes to the advancement of knowledge representation in e-Government****,*** *showcasing the potential of KGE to support intelligent decision-making, improve service delivery, and foster a more data-driven public administration.*

Keywords: *Knowledge Graph, Machine Learning, Knowledge Graph Embedding, e-Government, Artificial Intelligence*

1. Introduction

In an era of data-driven intelligence, structured knowledge representation plays a pivotal role in advancing artificial intelligence (AI) and machine learning (ML) applications. Knowledge Graphs (KGs) have emerged as an essential paradigm for encoding complex relationships between entities, enabling machines to reason, infer, and generate insights across diverse domains, including healthcare, finance, social networks, and particularly e-Government systems. A KG organizes information as a set of triples (head entity, relation, tail entity), forming a structured representation of real-world facts. e-Government is the use of Information and Communication Technology (ICT) within public policy, operations in public organizations, citizens engagement, and government services. e-Government is a unique domain, and attempts have been made to use KG and exploited this uniqueness in solving problems in data management, data transformation, and web service composition[1], [2], [3], [4] within the domain.

Knowledge Graph Embedding (KGE) are ML techniques that aim to transform symbolic knowledge into dense, low-dimensional vector spaces while preserving the structural and semantic properties of the original KG. These embeddings facilitate various downstream tasks, such as link prediction, entity classification, relation extraction, and recommendation systems. By leveraging representation learning, KGE methods enable the effective integration of heterogeneous data sources, improving knowledge discovery and decision-making. Traditional KGE approaches, such as TransE [5], TransH [6], and TransR[7], model relationships as translation operations in vector space, while more recent methods employ deep learning techniques, including graph neural networks (GNNs) and transformer-based architectures, to capture complex dependencies.

An emerging research area in the field of AI is neuro-symbolic AI, which seeks to combine the strengths of symbolic reasoning with the adaptability of neural networks [8]. Neuro-symbolic AI enhances representation learning by integrating logical constraints, rule-based reasoning, and domain-specific ontologies with deep learning models. In the context of e-Government, this approach provides a means to ensure interpretability, improve reasoning over sparse KGs, and enable more reliable decision-making processes. By embedding structured knowledge with logical constraints, neuro-symbolic AI can help address key challenges such as knowledge incompleteness, generalization across governmental domains, and the integration of multimodal data sources.

Despite significant progress, several challenges remain in KGE research for e-Government applications. Scalability remains a key concern, as e-Government KGs often contain vast amounts of data related to public services, legal frameworks, and administrative processes. Additionally, ensuring interpretability and generalization across different governmental domains continues to be an open problem. The integration of multimodal data sources, temporal reasoning, and knowledge-aware pretraining in large-scale models presents promising yet complex research directions.

In this paper, we apply a KGE model – TransE, on an e-Government dataset, as a continuation of past works on application of KG in the e-Government domain, and as a path towards KG in e-Government in the neuro-symbolic research field.

2. Related Work

Several works have been carried out in various graph embedding techniques, and these works can be broadly classified into Matrix Factorization (MF)-based - [9], [10], [11], Graph Neural Networks (GNN)-based - [12], [13], [14], Autoencoder-based - [15], and KGE-based - [5], [6], [7].

Our work applies KGE to the e-Government domain using our unique context and dataset.

3. KG Embedding Approach

3.1 Dataset and Experiment Setup

The dataset used in this study is an RDF/OWL knowledge graph representing e-Government information. It was developed using Protégé, a widely used ontology editor, and contains entities and relationships structured in accordance with semantic web standards. This knowledge graph serves as the foundation for applying Knowledge Graph Embedding (KGE) techniques, specifically for link prediction and graph completion. Since real-world knowledge graphs often suffer from incompleteness, this study focuses on predicting missing links to enhance knowledge inference. No external ontologies were integrated into this work; instead, the dataset's built-in schema was used to structure relationships and entity representations. The RDF/OWL file was preprocessed using RDFLib, which facilitated the extraction of triples in the form (head entity, relation, tail entity). These triples were then converted into a format compatible with embedding models. Additionally, NetworkX was employed to visualize graph structures and gain insights into entity connectivity.

Experiments were conducted on a PC with an Intel Core i7 CPU @ 1.90GHz, 16GB RAM, running Windows 11 (64-bit OS). The implementation was developed using PyTorch for tensor computations, rather than deep learning, learning computations, PyKEEN for KGE model training, and supporting libraries such as RDFLib, NetworkX, Matplotlib, NumPy, TSNE, and Seaborn (SNS) for data manipulation, visualization, and dimensionality reduction.

3.2 Embedding Model and Training

For knowledge representation learning, we employed TransE, a simple yet efficient translational embedding model. TransE represents relationships as translation vectors in a continuous low-dimensional space, ensuring that the sum of the head entity and relation vectors approximates the tail entity vector. It was chosen for its scalability and ease of interpretability, making it a suitable choice for an initial exploration of KGE on e-Government data. At this stage, no neuro-symbolic AI techniques or additional logical constraints were incorporated. Additionally, no hyperparameter tuning was performed, as the primary focus was on applying and evaluating TransE embeddings within the e-Government domain.

The dataset was split into 80% training, 10% validation, and 10% test sets to ensure a fair evaluation. The TransE model was trained using the PyKEEN framework, which provides an efficient implementation of various KGE models. Training parameters followed the default configurations provided by PyKEEN, given that hyperparameter optimization was not conducted in this phase. For evaluation, we analyzed standard KGE performance metrics, including Mean Rank (MR), Mean Reciprocal Rank (MRR), and Hits@k (k = {1, 3, 10}), to assess link prediction accuracy. However, the primary goal of this study was not performance benchmarking but rather the application of TransE embeddings on an e-Government knowledge graph to gain insights into its structural properties and link predictability.

4. Result and Discussions

Two distinct colors – blue and red, are used to represent the head and tail entities respectively, in the dataset as shown in figure 1. The blue nodes (inner cluster) are more densely packed, indicating entities that serve as sources (heads) in many relationships. The tail nodes seem to form a boundary around the graph (outer circle), indicating entities that are predominantly targets(tails) of the relationships. The directed edges (arrows) represent relationships between entities, showing the flow from head to tail entities. The numerical edge labels are the relationship type associated with each edge. The repetition of certain numbers indicates common relationships between entities. The densely connected core (blue nodes) has high connectivity, indicating that these entities are highly related. The peripheral nodes (red outer circle) have fewer connections and appear more isolated, indicating interaction mostly with the core entities. Certain blue nodes are acting as hubs that connect multiple relationships, indicating key entities with many interactions.

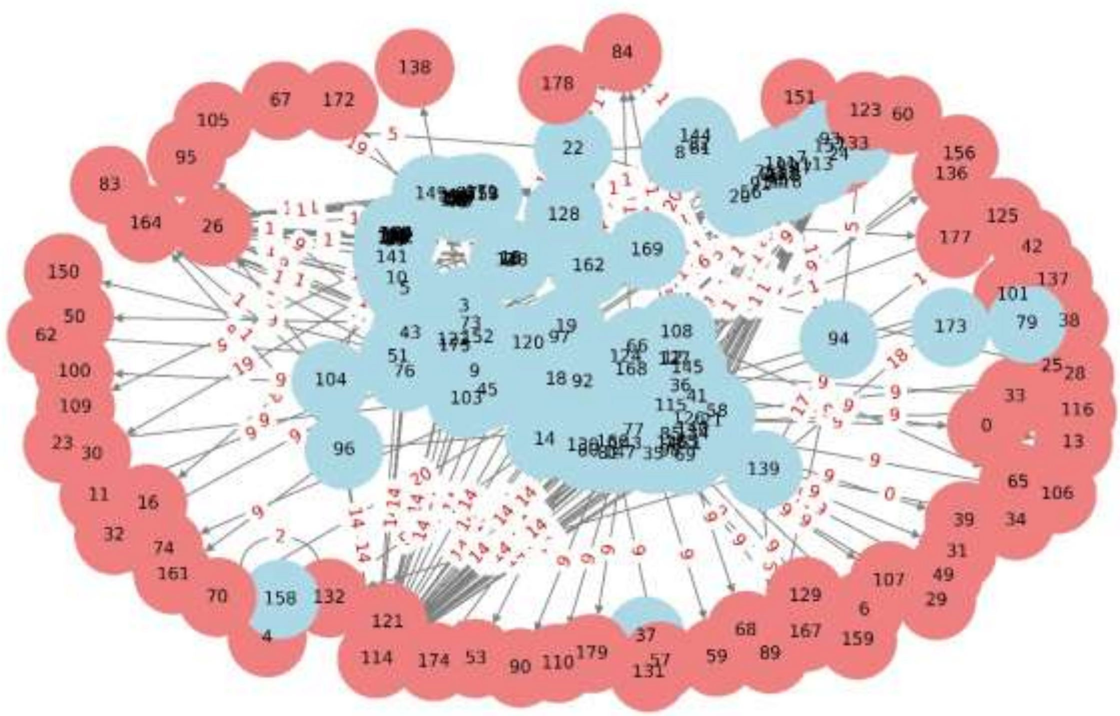

Figure 1. Graph Visualization of the e-Government dataset with Different Node Colors for Heads and Tails

The output of the evaluation metrics of our trained TransE model in predicting KG facts, is shown in Table 1, including the first 5 and last 5 items. The table contain the following columns:

Side – Whether the metric is computed for the head, tail, or both entities in the triples.

Rank_type – The ranking strategy used (optimistic, realistic, pessimistic)

Metric – The specific evaluation metric (e.g., variance, hits@k, adjusted_hits_at_k)

Value – The numerical value of the metric

The variance is quite high (~2500-2700 across conditions), indicating that the model's performance varies significantly depending on sample. Hits@K performance varies depending on head vs. tail predictions, indicating that the model might be better at predicting tails than head. Negative "adjusted_hits_at_k" for head, pessimistic, indicates poor performance in worst-case scenarios when predicting missing heads.

Looking predictions output, we can investigate where performance drops. For example, if tail predictions are better than head predictions, we can consider balancing entity representations during training. High variance suggests inconsistent generalization, and we can increase the embedding size or tune hyperparameters to improve embeddings. Lastly, if the adjusted Hits@K score is negative or too low, it might indicate an issue with dataset bias or the evaluation method.

Table 1. KG Embedding Evaluation Metrics

	Side	Rank_type	Metric	Value
0	Head	Optimistic	Variance	2259.678367
1	Tail	Optimistic	Variance	2528.191020
2	Both	Optimistic	Variance	2671.871224
3	Head	Realistic	Variance	2259.678467
4	Tail	Realistic	Variance	2528.191406
220	Tail	Realistic	Adjusted_hits_at_k	0.213364
221	Both	Realistic	Adjusted_hits_at_k	0.088722
222	Head	Pessimistic	Adjusted_hits_at_k	-0.036918
223	Tail	Pessimistic	Adjusted_hits_at_k	0.213364
224	Both	Pessimistic	Adjusted_hits_at_k	0.088722

KG embeddings can be visualized using the t-SNE visualization tool, as shown in figure 2. The plot represents a 2D projection of high-dimensional KG embeddings using t-SNE. Each point corresponds to an entity in our dataset, and its position reflects relationships learned during training. Entities of the same type (color-coded) tend to cluster together, suggesting that similar entities have similar embeddings.

The legend shows different entities types (e.g., Location, Person, Citizen, Organization, Document). Some types are well separated, while others are intermingled. Densely clustered points indicate strong structural similarity, and scattered points indicate entities that have weaker semantic relationships. Entities that are not well separated may indicate overlapping roles in the KG, or suboptimal embedding training (e.g., insufficient training epochs or inappropriate hyperparameters).

Closer points represent entities with similar roles or meanings. Farther apart points represent entities with fewer shared properties in the KG. Adjusting t-SNE parameters like perplexity could help in making expected clusters appear clearly.

The non-deterministic nature of t-SNE means different runs might yield slightly different results, and it captures local structures well but might distort global relationships. These are some limitations of t-SNE, as a visualization tool.

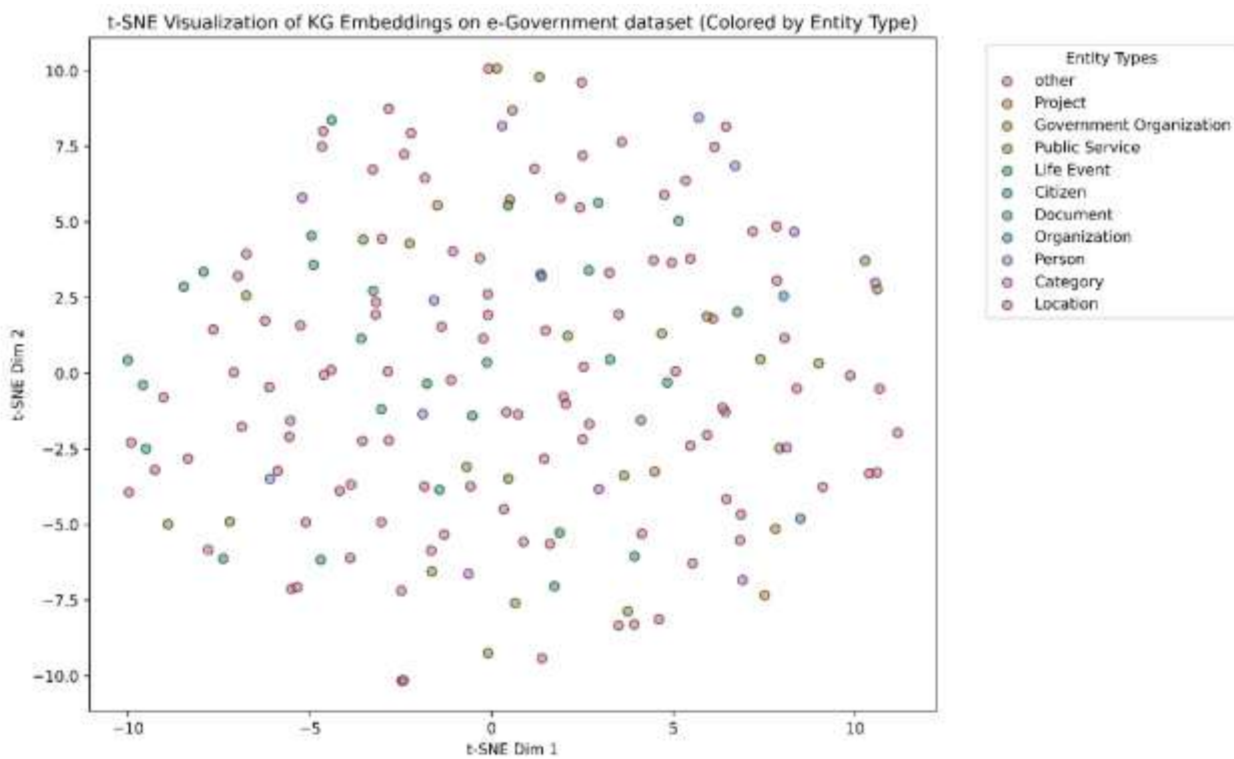

Figure 2. Visualization of KG Embeddings on e-Government Dataset

The result of a tail entity prediction task using our trained TransE model on our e-Government KG is shown in figure 2. The model was asked to predict the most likely tail entity for a given (head, relation, ?) triple. The result consists of a TargetPredictions object, shown in the table with the following columns:

Tail_id: The numerical ID of the predicted tail entity

Score: The score assigned by the model to each prediction (lower scores are better in TransE)

Tail_label: The actual name (URI) of the predicted tail entity

Additionally, the model predicting the tail entity given a head entity and relation.

Table 2. Predictions of Tail for head entity = "urn:absolute:eGov.RiversState#Aziboala_Nwuche", showing the first 5, and last 5 predictions.

Target Predictions (df=	Tail ID	Score	Tail Label
81	81	-7.056051	urn:absolute:eGov.RiversState#Aziboala_Nwuche
66	66	-7.585266	http://www.w3.org/2002/07/owl#NameIndividual
103	103	-8.310295	Urn:absolute:eGov.RiversState#Government_Instituti on
42	42	-8.364018	2023-11-17T12:03:44+00:00
121	121	-8.445612	Urn:absolute:eGov.RiversState#Non-Indigene
47	47	-11.384985	2023-11-17T13:36:09+00:00
16	16	-11.453408	2023-08-27T17:09:28+00:00
18	18	-11.470187	2023-08-27T17:15:42+00:00
146	146	-11.796174	urn:absolute:eGov.RiversState#Sanclin_Hospital
106	106	-11.900414	urn:absolute:eGov.RiversState#Health_Care_Provider

The TransE model assigns scores to each possible tail entity. Lower scores indicate stronger confidence in the prediction, and the predicted tail entity with the lowest score is considered the most plausible. The top-ranked entity is Aziboala_Nwuche (ID 81) with the lowest score (-0.7056); the model predicts that this entity is the most plausible as the missing tail. Other entities have slightly higher (less negative) scores; these entities are less likely but still plausible. Entities with timestamps are predicted as tails; this indicates that temporal information is embedded in the KG. Some standard RDF concepts (owl:NameIndividual) appear, and filtering out such generic entities might improve the predictions. On the whole, the output shows the predicted tail entities for a given query, ranked by plausibility. The most plausible tail (Aziboala_Nwuche) has the lowest score (-0.7056), but all scores are negative, meaning even the best prediction is not highly confident. The presence of timestamps and generic RDF entities may indicate areas for improvement in the model.

5. Conclusions

This paper presents an initial exploration of Knowledge Graph Embedding (KGE) on an e-Government knowledge graph using TransE. Future work will focus on enhancing model performance through hyperparameter tuning, incorporating neuro-symbolic AI techniques, and integrating external ontologies to improve semantic reasoning. The findings from this work contribute to understanding how embedding models can be applied to structured governmental data, enabling better knowledge inference and decision support systems.

6. References

[1] F. Orji, N. Nwiabu, B. Okoni, and O. Taylor, "A Knowledge Graph Model for e-Government," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 4, Apr. 2024.

[2] F. Orji, N. Nwiabu, O. Bennett, and O. Taylor, "A Knowledge Graph Data Transformation Model for e-Government," *E-Gov.*, vol. 2, no. 3, p. 4, 2024.

[3] F. Orji, N. Nwiabu, O. Bennett, and O. Taylor, "Artificial Intelligence in e-Government: A Computational Perspective Review".

[4] F. Orji, N. Nwiabu, B. Okoni, and O. Taylor, "A Knowledge Graph Service Composition Model for e-Government," *Int. Res. J. Comput. Sci.*, vol. 11, no. 4, pp. 145–152, 2024.

[5] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[6] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the AAAI conference on artificial intelligence*, 2014.

[7] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proceedings of the AAAI conference on artificial intelligence*, 2015.

[8] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, and L. Zhou, "Neuro-symbolic approaches in artificial intelligence," *Natl. Sci. Rev.*, vol. 9, no. 6, p. nwac035, Jun. 2022, doi: 10.1093/nsr/nwac035.

[9] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.

[10] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[11] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *ArXiv Prepr. ArXiv160902907*, 2016.

[13] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[14] P. Velickovic *et al.*, “Graph attention networks,” *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.
[15] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.