Towards Transparent Knowledge Graphs: A Position on Explainability in Link Prediction

Vidhya Kamakshi[1], Chandramani Chaudhary[1]

[1]Department of Computer Science & Engineering, National Institute of Technology Calicut, Kerala - 673601, India.

Abstract

Knowledge Graphs (KGs) have improved structured knowledge representation by encoding real-world entities and their relationships, enabling multi-hop reasoning for answering complex queries. However, state-of-the-art deep learning models applied to KGs lack interpretability, creating a challenge in understanding their decision-making processes. This paper presents an idea to integrate Explainable AI (XAI) techniques with knowledge graph embeddings to enhance transparency in link prediction models. We employ SHAP (SHapley Additive exPlanations), a game-theoretic approach, to quantify the influence of individual entities in predictions. Furthermore, we introduce an explanation-driven training framework that aligns model predictions with the underlying KG structure. By incorporating an explainability-aware loss function, our approach may provide high-quality link predictions and human-understandable explanations. This research contributes to developing more transparent AI systems, fostering trust in real-world applications where interpretability is crucial.

Keywords

Explainable AI, Natural Language Processing, Knowledge Graphs, Interpretable AI, Trustworthy AI

1. Introduction

Natural Language Processing (NLP) is a sub-domain of Artificial Intelligence that deals with encoding world knowledge that is often expressed in Natural Languages like English, French, Hindi, etc., into a vector representation that can be processed by the models. Adequate representation is essential to enable the model to respond appropriately to the queries of the human users. The advent of deep models that base their prediction on the transformation of the sequences into semanticity preserving vector spaces has enhanced their capabilities in processing natural language human queries. Google introduced an intermediate representation called Knowledge Graph (KG) [1] that structured the semantic information available in the web. The graph has a set of real-world entities that are the nodes, and the relationships between these entities are encoded in its directed edges. This gave a novel perspective to processing natural language queries through a multi-hop traversal on the knowledge graph to extract related triples of the form (head entity, relationship, tail entity) that enables the model to respond to natural language queries. For instance, a query "Where is the captain of the Indian Cricket team born?" is successfully retrieved following multiple hops, retrieving triples (c, Captain, Indian Cricket Team), (c, Birth Place, p). The deep models that offer state-of-the-art performances bring in a novel problem of opacity, rendering the working mechanism of these underlying models uninterpretable to the end users [2].

The need for interpretability is increasing following the mandates from legal frameworks [3] that facilitate the user to know the rationale behind the decision of an AI model concerning the user. Eliciting explanations is necessary to identify biases [4], thereby assessing the suitability of deploying an AI model for real-world applications. Explanations can help spot the erroneous facts employed by the AI model, thereby guiding ways to correct these errors [5, 6] to inculcate the right rationale into the model.

In this paper, we explore the integration of explainable AI (XAI) techniques with knowledge graphs, addressing the need for transparency in link prediction models. Our approach leverages knowledge

1st International Workshop on Explainable AI and Knowledge Graphs (XAI+KG), JUNE 1-5, 2025, co-located with ESWC 2025, Portoroz, Slovenia

[†] The authors contributed equally.

vidhyakamakshi@nitc.ac.in (V. Kamakshi); chandramanic@nitc.ac.in (C. Chaudhary)
0000-0001-7588-6318 (V. Kamakshi); 0009-0006-3497-1309 (C. Chaudhary)

graph embeddings [7, 8, 9, 10], to learn structured representations of entities and relations. Additionally, we incorporate SHAP (SHapley Additive exPlanations) [11], a game-theoretic method, to quantify the influence of individual entities in the prediction process. By introducing explanation-driven training, we enforce that our model efficiently leverages the underlying KG structure. The proposed framework improves optimal traversal, thus exhibiting increased interpretability.

2. Related Work

Knowledge Graphs (KGs) provide structured semantic representations and are central to many AI applications. Open KGs such as Freebase [12], DBpedia [13], and YAGO [14] have spurred research on KG embeddings for link prediction. Early methods, including translational models and semantic matching approaches such as tensor decomposition, project entities into vector spaces to infer missing links [1]. Recent deep learning approaches, such as Graph Convolutional Networks (GCN) [15], Graph Auto-Encoder Attention Networks, and Relational GCNs, integrate KG structure directly into end-to-end models [1]. Despite state-of-the-art successes exhibited by the deep NLP models, their opacity inhibiting the understanding of its rationale may prove detrimental if blindly employed in safety-critical applications [16]. This calls for developing tools and techniques to open up these accurate black boxes and investigate their working mechanisms.

Explainable AI (XAI) aim to demystify the black box models. These techniques can be broadly classified into antehoc or explainable by design approaches and posthoc approaches. Antehoc techniques inculcate the ability to explain the action a model takes from the design phase of the model. They are applied when a model is yet to be constructed and faithfulness is of utmost concern [17]. On the other hand posthoc techniques construct a simpler explainer that mimics the working mechanism of a black box model leaving it undisturbed. When a model is already deployed, posthoc techniques [18] are usually the desired mode of incorporating explanations into the model pipeline. There have been domain-specific and model-specific [19] techniques that have been proposed to extract explanations from the deployed models in a posthoc manner. Alternately XAI community has also proposed model-agnostic techniques [11, 20] that can be leveraged for any data modality and models. These techniques have been applied to various NLP tasks [2]. Transformer architectures [21] which leverage self attention mechanism, are designed to handle long range dependencies. There have been attempts to leverage these attention maps [22, 23] as an explanation to the model's working mechanism, which may spark debates in the research community [24] concerning their suitability to faithfully explain the black boxes.

An alternate way to incorporate explainability into the NLP models is to relate the rationale of the black box model with the knowledge encoded in the knowledge graph representations. While prominent works in the community [25, 26] explore the direction of leveraging and aligning NLP models with known knowledge encoded in the knowledge graph, this paper calls for an idea to leverage XAI techniques for tracing the path traversed by the model and reinforce the model to traverse optimal paths in the knowledge graph while performing link prediction. Rossi et al. [27], whose intent is close to ours, propose generating a local explanation by identifying necessary and sufficient entities that determine the prediction. Our proposed approach relies on Shapley values [11] with a strong game theoretic backing to globally rank the entities based on their influence in the prediction.

3. An Idea to Optimize Knowledge Graph Traversal using XAI

3.1. Knowledge Graph Representation

The knowledge graph (KG) is modeled as a labeled directed graph $G = (V, E)$, where V represents entities manifesting as nodes of the graph and E represents relations manifesting as the directed edges between the entites. The graph structure is used to learn node embeddings that capture semantic relationships between entities. Typically, KGs are represented with triplets, (h, r, t), where h is the head entity, t is the tail entity, and r is the relation between them.

3.2. Knowledge Graph Embedding Model

The proposal is flexible to accommodate any knowledge graph embedding models, such as ComplEx [7], TransE [8], DistMult [9], or RotatE [10]. For learning the KG embeddings, a margin-based ranking loss that refines the embeddings may be adopted, whose formulation is as follows:

$$\mathcal{L}_{emb} = \sum_{(h,r,t)\in\mathcal{P}} \sum_{(h',r,t')\in\mathcal{N}} \max(0, s(h',r,t') - s(h,r,t) + \lambda) \tag{1}$$

Here, s is the score function given by the embedding model, $\mathcal{P}$ is the set of positive triplets, $\mathcal{N}$ denotes the set of negative triplets, and λ denotes the tolerable margin that controls separation between the triplets of opposing polarity.

3.3. Link Prediction Model

A Graph Convolutional Network (GCN) [15] can be leveraged to predict the missing links in a KG by processing the learned entity and relation embeddings as a composition of non-linear activations applied on linearly combined features. This can be mathematically expressed as follows:

$$x' = \sigma(W_1 \cdot GCNConv(x))$$
$$x'' = W_2 \cdot GCNConv(x')$$

where W_1 and W_2 are learnable weight matrices, and σ is a non-linear activation function. The refined embeddings are projected onto the relation space (the projection is characterized by W_3), followed by the application of softmax function to predict the most likely relation type for a given entity pair:

$$\hat{y} = softmax(W_3 x'')$$

A categorical cross-entropy loss may be applied to ensure alignment between predicted ($\hat{y}$) and ground truth relation (y) defined as:

$$\mathcal{L}_{\text{pred}} = -\sum_{(h,r,t)\in\mathcal{P}} \sum_{j=1}^{|R|} y_j \log(\hat{y}_j) \tag{2}$$

where, $|R|$ is the total number of relation types, y_j is the one-hot encoded ground truth relation, $\hat{y}_j$ is the predicted probability for relation j.

3.4. Calculation of SHAP(v)

The computation of SHAP values [11] in Game theory to quantify the importance of each player in a game proceeds through simulations where a player is removed from the team (set) and the effective score of the team (subset) with the deletion is used to estimate the contribution of that player to the game. The translation of this phenomenon in the KG lingua is discussed in this section.

3.4.1. Model Input Representation

The SHAP explainer takes as input the node embeddings learned through the knowledge graph embedding model, which effectively captures the semantic relationships between entities. For a given entity pair (h, t), the embeddings corresponding to the head entity h and the tail entity t are extracted. For each entity pair (h, t), we select a subgraph that captures the local structural context of the KG. This subgraph is determined by extracting nodes within a predefined radius of both h and t. In cases where multiple shortest paths exist between h and t, our approach will detail one of the following strategies:

- Aggregation: Compute SHAP values for each shortest path separately and then aggregate (e.g., via averaging) the contributions.
- Selection: Use a heuristic (e.g., the path with the highest cumulative link prediction score) to select the most representative path.

3.4.2. Perturbation-Based Feature Importance

SHAP employs a perturbation-based approach to determine feature importance by systematically modifying input features, specifically the node embeddings, and analyzing their effect on the link prediction model. This is achieved by masking or removing different subsets of nodes within the selected subgraph to observe how these alterations influence the model's predictions. The trained link predictor is then used to recompute predictions for each perturbed version of the input, allowing for the quantification of the contribution of individual nodes to the final prediction. This process helps in understanding how different nodes in the knowledge graph influence the model's decision-making.

3.4.3. Shapley Value Estimation

The SHAP framework approximates Shapley values, which quantify the contribution of each node to the final link prediction decision. Let V denote the set of all nodes in the knowledge graph, and let $S \subseteq V \setminus \{v\}$ represent a subset of nodes excluding node v. The contribution of each node v to the link prediction task is computed using the Shapley value formula:

$$SHAP(v) = \sum_{S \subseteq V \setminus \{v\}} \frac{|S|!(|V| - |S| - 1)!}{|V|!} (f(V) - f(S))$$

where $f(S)$ denotes the link predictor's score when only the nodes in subset S are included, and $f(V)$ denotes the link predictor's score when all the nodes are used. The term $f(V) - f(S)$ captures the marginal impact of adding node v to subset S. The weighting factor $\frac{|S|!(|V|-|S|-1)!}{|V|!}$ ensures a fair distribution of contributions across all possible subsets. By systematically evaluating the marginal contribution of each node across different subsets, this method provides a robust measure of the importance of individual nodes in influencing the link prediction outcomes.

Since computing exact Shapley values is computationally expensive [28], we approximate them using Kernel SHAP or Deep SHAP, which efficiently estimates contributions using a smaller subset of perturbations [29]. The Shapley values signifying the extent of influence of a node are normalized to facilitate comparability across different entity pairs.

3.5. Explainability-Driven Training Framework

To leverage the explanations for iterative model improvement a score that assesses the explanations (i.e. contribution scores of each node) with respect to a shortest path P between head entity h and tail entity t in the ground-truth KG is formulated as follows:

$$S_{exp} = \frac{1}{|P|} \sum_{v \in P} SHAP(v) \tag{3}$$

A lower score indicates poor alignment between predictions and the actual KG structure.

3.6. Loss Function

A composite loss function that balances classification accuracy, explainability, and embedding optimization may be formulated as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{emb} + \beta \cdot \mathcal{L}_{pred} + \gamma \cdot \mathcal{L}_{exp}$$

where, $\mathcal{L}_{emb}$ as formulated in equation 1 ensures learning high-quality KG embeddings, $\mathcal{L}_{pred}$ is the cross-entropy loss for relation prediction (softmax output) as formulated in equation 2, $\mathcal{L}_{exp} = 1 - S_{exp}$ (formulated in equation 3) penalizes traversing sub-optimal paths, and α, β, and γ control the trade-off between embedding optimization, accuracy, and interpretability .

By integrating explainability into the learning process, the model not only predicts links accurately but also provides interpretable insights into its decisions. This approach ensures that the learned embeddings and model predictions remain aligned with the intrinsic structure of the knowledge graph.

4. Summary

The paper reviews the scientific literature and identifies a symbiotic relationship between Knowledge Graphs and Explainable AI research communities. A framework to incorporate explainability techniques as a guiding mechanism towards steering the NLP model to faithfully traverse through optimal paths in the knowledge graph is suggested. An illustration using commonly used knowledge graph embedding and link prediction model with their corresponding mathematical formulations has been presented to encourage the research community to investigate this incorporation. Modification of these techniques with other state-of-the-art algorithms is an open arena that may yield novel insights.

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates. The authors place their heartfelt gratitude to the resources provided by the institute they are affiliated and their department which is sponsored by the prestigious DST-FIST Initiative of the Government of India. We sincerely thank the reviewers whose constructive feedback has helped shape the camera ready draft of the paper.

Declaration on Generative AI

During the preparation of this work, the authors used Chat-GPT-3.5 for Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] M. Wang, L. Qiu, X. Wang, A survey on knowledge graph embeddings for link prediction, Symmetry 13 (2021).

[2] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, F. A. Batarseh, Rationalization for explainable nlp: a survey, Frontiers in Artificial Intelligence 6 (2023) 1225093.

[3] Council of European Union, 2018 reform of eu data protection rules, 2018. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[4] H. Ahsan, A. S. Sharma, S. Amir, D. Bau, B. C. Wallace, Elucidating mechanisms of demographic bias in llms for healthcare, arXiv preprint arXiv:2502.13319 (2025).

[5] S. Cheng, N. Zhang, B. Tian, X. Chen, Q. Liu, H. Chen, Editing language model-based knowledge graph embeddings, in: Proceedings of the AAAI Conference on Artificial Intelligence, 16, 2024, pp. 17835–17843.

[6] H. Gu, K. Zhou, X. Han, N. Liu, R. Wang, X. Wang, Pokemqa: Programmable knowledge editing for multi-hop question answering, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 8069–8083.

[7] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: International Conference on Machine Learning, PMLR, 2016, pp. 2071–2080.

[8] J. Weston, A. Bordes, O. Yakhnenko, N. Usunier, Connecting language and knowledge bases with embedding models for relation extraction, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1366–1371.

[9] B. Yang, S. W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, in: International Conference on Learning Representations, 2015.

[10] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, in: International Conference on Learning Representations, 2019.

[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30 (2017).

[12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2008, pp. 1247–1250.

[13] D. Ringler, H. Paulheim, One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co., in: KI 2017: Advances in Artificial Intelligence: 40th Annual German Conference on AI, Dortmund, Germany, September 25–29, 2017, Proceedings 40, Springer, 2017, pp. 366–372.

[14] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 697–706.

[15] H. Zhang, G. Lu, M. Zhan, B. Zhang, Semi-supervised classification of graph convolutional networks with laplacian rank constraints, Neural Processing Letters (2022) 1–12.

[16] B. Koopman, G. Zuccon, Dr chatgpt tell me what i want to hear: How different prompts impact health answer correctness, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023, pp. 15012–15022.

[17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[18] V. Kamakshi, N. C Krishnan, Sce: Shared concept extractor to explain a cnn's classification dynamics, in: Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD), 2024, pp. 109–117.

[19] N. Mylonas, I. Mollas, G. Tsoumakas, An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification, Data Mining and Knowledge Discovery 38 (2024) 128–153.

[20] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[22] E. A. Shams, J. Carson-Berndsen, Attention to phonetics: A visually informed explanation of speech transformers, in: International Conference on Text, Speech, and Dialogue, Springer, 2024, pp. 81–93.

[23] E. A. Shams, I. Gessinger, J. Carson-Berndsen, Uncovering syllable constituents in the self-attention-based speech representations of whisper, in: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, 2024, pp. 238–247.

[24] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, P. Watrin, Is attention explanation? an introduction to the debate, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 3889–3900.

[25] A. Füßl, V. Nissen, Interpretability of knowledge graph-based explainable process analysis, in: 2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), IEEE, 2022, pp. 9–17.

[26] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 1–33.

[27] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining link prediction systems based on knowledge graph embeddings, in: Proceedings of the International Conference on Management of Data, 2022, pp. 2062–2075.

[28] M. Arenas, P. Barceló, L. Bertossi, M. Monet, The tractability of shap-score-based explanations for classification over deterministic and decomposable boolean circuits, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 6670–6678.

[29] S. Akkas, A. Azad, Gnnshap: Scalable and accurate gnn explanation using shapley values, in: Proceedings of the ACM Web Conference, 2024, pp. 827–838.