SPHOTA: Knowledge Graph Structure Prediction with a Hybrid Orientation of Textual Alignment using K-BERT

R. P. Bharath Chand[1], Sanju Tiwari[2]

[1]Department of Futures Studies, University of Kerala, India
[2]Sharda University, Delhi-NCR, India

Abstract

Link prediction in domain-specific knowledge graphs presents unique challenges that require both structured data representation and contextual understanding. Existing approaches typically focus either on topological embeddings or language models, but rarely integrate both in a unified framework. This work addresses that gap by proposing a hybrid approach that integrates Knowledge Graph Embedding techniques with Language Models to enhance link prediction performance. Our aim is to enhance the link prediction performance in the context of the Second International Biochemical Knowledge Extraction Challenge. The methodology involves evaluating knowledge integration strategy of K-BERT, combined with a regularization model of EPHEN method, applied to the NuBBE dataset. The approach focuses on designing an optimal knowledge integration pipeline to improve the predictive accuracy. The study anticipates that this integrated framework will advance domain-relevant link prediction and set a foundation for the future research into tighter KG–LLM coupling strategies.

Keywords

Knowledge Injection, K-BERT, Knowledge Graph, Knowledge Graph Embedding, Large Language Model, Link Prediction, Bio-Chemical

1. Introduction

Techniques used for Knowledge Graph Embedding include topology-based methods and language model-based embedding approaches, each with distinct advantages as shown in Table 1. Topology-based embeddings offer structured data representation but lack the contextual richness of language models. Conversely, language models provide context awareness, which structured knowledge graphs often miss—sometimes leading to erroneous predictions. Therefore, both approaches are complementary and can enhance tasks such as link prediction when used together.

Traditional embedding methods predominantly focus on network structure, but EPHEN [1] model takes a different approach by leveraging a language model to propagate embeddings. It integrates both event descriptions and their intricate relationships into a low-dimensional vector space, allowing for smooth and adaptive embedding updates [2]. This model is currently the only one of its kind applied to the Knowledge Extraction task in the Biochemical Knowledge Extraction Challenge [3] using NuBBE dataset [4].

Biochemical compounds are central to pharmaceutical innovation, yet their discovery remains hindered by the fragmented and unstructured nature of textual data in academic publications. Predicting chemical compounds or other related information from literature requires a nuanced understanding of contextual relationships between factors like the plant species, compound name, bioactivity, collection site, isolation type, etc. By leveraging unsupervised Knowledge Graph Embedding techniques and Language Models, we can infer the missing information and predict novel associations from semantic and structural patterns in existing data. In this context, our proposed method is for investigating the feasibility associated with a hybrid model and apply these insights to the BiKE challenge [3] to achieve improved performance through necessary modifications.

Table 1 outlines how Language Models and Knowledge Graphs offer distinct but complementary strengths when applied to link prediction tasks. Language Models excel in semantic understanding and

BiKE'25: Second International Biochemical Knowledge Extraction Challenge, June 1–2, 2025, co-located with Extended Semantic Web Conference (ESWC), Portorož, Slovenia.

can suggest plausible links by leveraging the context. However, they often operate as black boxes with limited explainability and are prone to hallucinating links that do not exist. In contrast, Knowledge Graphs use structured, graph-based approaches such as embeddings or path based methods to predict links. They tend to be less prone to noise or hallucination, especially when carefully curated.

Despite these strengths, KGs typically lack contextual awareness unless explicitly modeled. This is where Language Models can provide added value by incorporating real world knowledge and contextual information. Conversely, KGs can help validate or ground the predictions made by the Language Models, making the combined use of both approaches particularly powerful for improving the accuracy and trustworthiness of link prediction tasks.

Aspects	**Language Models**	**Knowledge Graphs**	**Complementarity**
Link Prediction	Predict links based on semantic understanding and context from training data.	Predict links using graph-based approaches (e.g., embeddings, path-based methods).	Language Models can suggest plausible links beyond existing graph patterns; KGs ground those suggestions in facts.
Explainability	Low – often black-box predictions	High – based on graph paths, schema, ontologies	KGs can help explain or validate the Language Models-predicted links through graph paths
Context Awareness	Strong. Can tell X is likely related to Y because of a contextual specification.	Weak, unless explicitly modeled	Language Models complement KGs with real-world contextual reasoning
Noise / Hallucination	Prone to hallucination (non-existent links)	Low if curated properly	KG can act as a filter to validate Language Models-predicted links

Table 1
Complementarity of Language Models and Knowledge Graphs in Link Prediction

2. Approach

This section outlines our approach to enhancing link prediction in a heterogeneous Knowledge Graph. The main strategy leverages recent advances in knowledge injection, embedding propagation, and similarity-based retrieval to unify semantic and topological signals within the graph, enabling more accurate and interpretable predictions.

2.1. Data Input

Dataset is derived from the NuBBE DB [4] natural product database, which was constructed through manual curation of over 2,000 peer-reviewed scientific articles by domain experts. These articles were annotated with key biochemical properties including compound name, bioactivity, species of origin, collection site, and isolation type. The dataset was then transformed into a knowledge graph (KG), where each scientific paper is represented by a central node (using its DOI), linked to various extracted properties as well as to topic nodes derived using BERTopic [5], a transformer-based topic modeling technique. The NatUKE benchmark paper [3] evaluates multiple unsupervised graph embedding techniques such as DeepWalk, Node2Vec, Metapath2Vec, and EPHEN for their effectiveness in completing such KGs via similarity-based link prediction, assessing how well each model predicts missing properties of natural products.

2.2. Knowledge Injection

Several methodologies exist for integrating KG content into language models, with the majority being Retrieval-Augmented Generation (RAG) based. RAG methods operate by modifying the input rather than altering the internal structure of the Language Modle. The Language Model remains intact, while RAG retrieves external knowledge and incorporates it into the input prior to processing. This renders RAG a non-invasive integration strategy. However, if RAG fails to retrieve pertinent documents that facilitate prediction, the model is unable to infer missing links independently, particularly when the relevant knowledge is distributed across multiple documents that exceed the token limit. Consequently, the efficacy of prediction is limited with the quality of knowledge retrieval prior to the Language Model processing.

Building on prior research, Cadeddu et al. [6] conducted a comprehensive analysis of different strategies for injecting structured knowledge into transformer-based language models. Their work focused on classification tasks; however, the intention in our work is to repurpose these strategies for link prediction tasks within hybrid models. All experiments in their study were conducted using BERT [7] as the base model. Nevertheless, Cadeddu et al. [6] emphasized the importance of exploring the applicability of these knowledge injection techniques across other contemporary large language models, such as LLaMA 2 [8]. Here we are considering only the BERT Language Model for this work using the K-BERT [9] approach. The following table provides an overview about the different other approaches. The combination of this can further be used for testing for its feasibility in the future. Table 2 is divided into two subtables: one for the Knowledge injection approaches mentioned in the analysis and one is the possible potential Language Models which is yet to be tested for [6]. Here, for our purpose of text embedding from the collected papers we chose K-BERT [9] method and added those embeddings as an attribute for each node of the knowledge graph in a similar way to how the EPHEN method used the SentenceBERT [10].

(a) Knowledge Integration Methods

Model	Method
DTI	Direct Text Injection
PT	Pretraining on triple text
MLP	Using Multilayer Perceptron
K-BERT	**Integrating triples**

(b) Language Models to be applied with

Model	Status
BERT	**Tested**
LLaMA 2	Not yet
GPT-2	Not Yet
GPT-J	Not Yet

Table 2
Overview of Knowledge Integration Methods; from which the K-BERT being the approach we used for this work

2.3. Regularization

We adopted the regularization-based embedding propagation method proposed by do Carmo and Marcacini [1], designed for heterogeneous information networks (HINs) with both textual and non textual nodes. This approach employs contextual text embeddings generated using K-BERT for nodes containing textual data, and propagates these embeddings across the network using a regularization function. Specifically, it minimizes the distance between connected node embeddings while also preserving the original semantic representation of text nodes through a tunable regularization term. This allowed us to unify the representation space across different types of entities in our dataset, thereby enabling improved downstream learning and inference tasks. The ability to incorporate both topological and semantic information was effective in enhancing the performance and interpretability of our model.

2.4. Link Prediction

After node embeddings are learned either through contextual embedding propagation (as in EPHEN and our proposed method) or structural path-based embedding (as in DeepWalk, Node2Vec and Metap-

ath2Vec) link prediction is performed using a K-nearest neighbors (KNN) approach in the embedding space. Each node in the knowledge graph is represented as a dense vector that encodes its semantics and structure. To predict links or retrieve associated properties (like species or bioactivity), the system computes cosine similarity between the embedding of the query node and all other candidate nodes of the target type. These candidates are then ranked based on similarity, and the top-k most similar nodes are selected as predictions. This is essentially a nearest-neighbor retrieval operation in a high-dimensional vector space, where similarity corresponds to semantic and relational closeness in the original heterogeneous information network (Knowledge Graph).

This KNN-based ranking mechanism enables evaluation using standard metrics such as hits@k (whether the correct link is among the top-k predictions) or MRR@k (Mean Reciprocal Rank), which emphasizes the rank position of the first correct prediction.

3. Evaluation

The key aspect of the evaluation strategy is deliberately disturbing the Knowledge Graph during testing by removing the links to the ground-truth property values for selected test nodes—specifically. This simulates a real-world knowledge extraction scenario where information is incomplete or missing. The model's task is to restore these missing links by ranking candidate nodes based on their embedding similarity to the test node. This setup is particularly meaningful for evaluating unsupervised graph embeddings, as it tests the model's ability to infer correct relationships purely from the remaining structure and contextual cues (like co-occurring topics or known relationships in the graph). The use of hit@k here measures how often the correct, previously removed value reappears within the top-k predicted candidates, thereby reflecting the model's capacity for knowledge graph completion. This approach enables a rigorous and realistic evaluation of how well each embedding method performs.

The evaluation was perfromed using the official BiKE challenge benchmark NatUKE [3] and the results are shown in the Table 3. It is a comparison with the final result of our proposed approach with the benchmark results given from official BiKE challenge. The table provides a comparative evaluation of different embedding and extraction methods DeepWalk, Node2Vec, Metapath2Vec, EPHEN, and the proposed K-BERT regularization method across five information extraction tasks for compound name (C), bioactivity (B), species (S), collection site (L), and isolation type (T). The metric used is the hit@k score for k = 50, 5, 20, 1, with the best results for each task and method in bold.

The proposed method of K-BERT with regularization approach consistently outperformed the baseline models across four extraction tasks - bioactivity, species, collection site, and isolation type. In particular, the most notable gains were observed in bioactivity and collection site prediction, where the method achieved hits@k scores significantly higher than all the baseline models. In the case of species and isolation type, the proposed method does not uniformly outperform all baselines across every evaluation steps. However it still delivers competitive results in the first evaluation stage for the species and first three evaluation stages for the isolation type. Overall, these findings suggest that the integration of contextual embeddings via K-BERT, along with embedding propagation through regularization, provides a richer and more discriminative representation of heterogeneous node types.

The open repository containing instructions on how to verify the claimed results is publicly available from this GitHub link https://github.com/bharathchand10/BiKE-SPHOTA

4. Related Work

Several recent works have aimed to improve the extraction of biochemical knowledge from scientific literature, particularly within the context of the NatUKE benchmark. One such effort, 'Leveraging ChatGPT API for Enhanced Data Preprocessing in NatUKE' [11], integrates OpenAI's ChatGPT into the data preprocessing pipeline to extract structured information, including compound names, bioactivity, species, collection site, and isolation type from the PDFs. This use of a general purpose language model serves to enrich the quality of input data before downstream processing.

Table 3
Results table for extracting: compound name (C), bioactivity (B), specie (S), collection site (L), and isolation type (T). Performance metric with the `hits@k` and k is respectively: 50, 5, 50, 20, and 1. The best results for each extraction are bold.

Property	DeepWalk			
	1st	**2nd**	**3rd**	**4th**
C	0.08	0.00	0.00	0.00
B	0.41	0.12	0.10	0.07
S	0.37	0.24	0.27	0.25
L	0.56	0.41	0.38	0.29
T	0.25	0.14	0.14	0.09

Property	Node2Vec			
	1st	**2nd**	**3rd**	**4th**
C	0.08	0.00	0.00	0.00
B	0.41	0.07	0.03	0.03
S	0.36	0.22	0.25	0.24
L	0.57	0.36	0.28	0.23
T	0.10	0.07	0.05	0.01

Property	Metapath2Vec			
	1st	**2nd**	**3rd**	**4th**
C	**0.10**	**0.08**	**0.09**	**0.20**
B	0.27	0.17	0.13	0.12
S	0.40	**0.41**	**0.42**	**0.44**
L	0.40	0.42	0.42	0.40
T	0.28	0.22	0.19	0.19

Property	EPHEN			
	1st	**2nd**	**3rd**	**4th**
C	0.09	0.02	0.03	0.04
B	0.55	0.57	0.60	0.64
S	0.36	0.24	0.29	0.30
L	0.53	0.52	0.55	0.55
T	0.71	0.66	0.75	**0.75**

Property	K-BERT regularization			
	1st	**2nd**	**3rd**	**4th**
C	0.09	0.00	0.00	0.00
B	**0.59**	**0.65**	**0.71**	**0.87**
S	**0.44**	0.29	0.32	0.32
L	**0.67**	**0.70**	**0.69**	**0.74**
T	**0.78**	**0.78**	**0.76**	0.71

Another contribution, 'Improving Natural Product Automatic Extraction With Named Entity Recognition' [12], enhances the traditional NatUKE pipeline by incorporating Named Entity Recognition (NER). Instead of basic token slicing, this approach extracts entire sentences that include target entities, thus preserving contextual integrity and improving entity-level accuracy in extraction.

The work 'Enhancing Biochemical Extraction with BFS-driven Knowledge Graph Embedding approach' [2] introduces a hybrid method combining Breadth-First Search (BFS) traversal of a knowledge graph with Word2Vec-based embedding. By treating BFS-generated paths as sentences, the model learns enriched vector representations for graph nodes, improving the capture of relational patterns in the literature.

Building upon these ideas, our work further investigates the synergy of structural embeddings and language models for improved link prediction in biochemical knowledge graphs. The following sections present the main works we directly adapt and build upon in our proposed approach.

4.1. Embedding Propagation over Heterogeneous Information Networks (EPHEN)

We have used EPHEN [1] as the upstream model to inspire and modify our downstream implementation. EPHEN is designed to integrate contextual text embeddings with the structural information of heterogeneous networks. It first generates embeddings for nodes containing textual data using a language model like SBERT and then propagates these embeddings to non textual nodes through a graph based regularization framework. This regularization minimizes the distance between connected node embeddings while preserving the original semantic representation of textual nodes. The resulting unified latent space allows effective comparison and analysis of both textual and non textual entities. Building on this methodology, we adapted and deployed a downstream version of the model with K-BERT.

4.2. A comparative analysis of knowledge injection strategies for large language models in the scholarly domain

The method of using K-BERT [9] for text embedding is adapted from the paper "A Comparative Analysis of Knowledge Injection Strategies for Large Language Models in the Scholarly Domain" by Cadeddu et al. [6]. This paper presents a thorough evaluation of different approaches for injecting structured knowledge into transformer models, particularly for scientific article classification. Among the strategies examined, K-BERT stands out for its sophisticated mechanism of knowledge injection through triple augmentation—appending relevant triples from a knowledge graph directly into the input text. K-BERT enhances token representations by constructing a sentence tree, which expands the sentence structure without overwhelming it, and uses a visible matrix to selectively control which tokens are allowed to influence each other during attention calculation. This ensures that only the pertinent triples affect the corresponding entities in the original sentence, preserving contextual integrity. The combination of the embedding layer, seeing layer, and mask self-attention mechanism allows K-BERT to learn rich, contextualised embeddings that incorporate both linguistic and structured knowledge efficiently.

5. Conclusion & Future Works

Among the evaluated methods, K-BERT with regularization emerges as one of the most consistent and accurate, leveraging both structural and semantic information to outperform others across various tasks. By leveraging the complementary strengths of structured topological embeddings and contextual language model representations, the proposed framework seeks to overcome the limitations inherent in using either approach independently. The development of an optimal knowledge integration architecture tailored to the chosen Language Model is expected to significantly improve predictive performance and application relevance.

Future work will focus on extending this framework through the design of novel, domain-sensitive loss functions that account for the varying impact of prediction errors in this specialized domain. Additionally, further exploration of advanced integration strategies such as graph augmented transformers and alternative architectures for tighter coupling between KGs and Language Models will also be pursued.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (https://chat.openai.com/) for assistance in refining academic language and enhancing grammatical accuracy.

References

[1] P. do Carmo, R. Marcacini, Embedding propagation over heterogeneous event networks for link prediction, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 4812–4821.

[2] B. Zope, S. Mishra, S. Tiwari, Enhancing biochemical extraction with bfs-driven knowledge graph embedding approach., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 235–243.

[3] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. S. e Silva, A. Pilon, Natuke: A benchmark for natural product knowledge extraction from academic literature, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), IEEE, 2023, pp. 199–203.

[4] A. C. Pilon, M. Valli, A. C. Dametto, M. E. F. Pinto, R. T. Freire, I. Castro-Gamboa, A. D. Andricopulo, V. S. Bolzani, Nubbedb: an updated database to uncover chemical and biological information from brazilian biodiversity, Scientific Reports 7 (2017) 7215.

[5] M. Grootendorst, Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics, Zenodo, Version v0 9 (2020).

[6] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Engineering Applications of Artificial Intelligence 133 (2024) 108166.
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[9] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, P. Wang, K-bert: Enabling language representation with knowledge graph, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 2901–2908.
[10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[11] P. Fröhlich, J. Gwozdz, M. Jooß, Leveraging chatgpt api for enhanced data preprocessing in natuke., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 244–255.
[12] S. Schmidt-Dichte, I. J. Mócsy, Improving natural product automatic extraction with named entity recognition., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 226–234.