An Empirical Study of the Consistency between Protocols for Evaluating Explanations of Predicted Links in Knowledge Graphs

Roberto Barile[1,*,†], Claudia d'Amato[1,2,*,†], Leonardo Santovito[1,*,†] and Nicola Fanizzi[1,2,*,†]

[1]Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Italy
[2]CILA, Università degli Studi di Bari Aldo Moro, Italy

Abstract

Since knowledge graphs are often incomplete, link prediction methods are adopted to predict missing facts. Although scalable embedding models are commonly used for this purpose, they lack comprehensibility, which may be crucial in several domains. Explanation methods address this issue by identifying pieces of knowledge that support the predicted facts. Regretfully, comparing quantitatively the resulting explanations is challenging because there are different protocols and no insights on their consistency when evaluating the same explanation method. Filling this important gap, we measure their consistency particularly as the correlation between the metrics resulting from evaluating the same explanation methods via different protocols. This requires evaluating the LP-X method CrossE in terms of a different protocol in addition to the ones introduced specifically for CrossE. We conduct experiments with different widely known knowledge graphs and embedding models. The outcomes suggest an overall consistency.

Keywords

Knowledge Graphs, Link Prediction, Explanation

1. Introduction

Knowledge Graphs (KGs) [1] are formal machine-processable representations of knowledge that conform to graph-based data models consisting of entities (nodes) and binary relations (edges). KGs deliver not only facts, but also intensional knowledge, which enables sound reasoning and is typically represented through ontologies. Despite their proven utility in academic and business [2, 3], KGs are often noisy and/or incomplete because the activities characterizing their life-cycle are often semi-automatic, incremental, and distributed [1]. *Link Prediction* (LP) methods aim at completing KGs by predicting missing facts and they mostly ground on *Knowledge Graph Embedding* (KGE) models that lead to competitive accuracy and scalability [4]. KGE models are representation learning solutions that encode the elements of a KG as low-dimensional vectors (embeddings), preserving their structural properties, that can be leveraged for tackling complex downstream tasks, such as LP, using efficient linear algebra operations. Despite such advantages, these models lack comprehensibility, i.e., are not traceable in terms of operations on symbolic/explicit knowledge. This problem hampers the use of LP via KGE models particularly in fields where it is paramount that stakeholders comprehend predictions before relying on them for making decisions with critical consequences. For example, the prediction of side effects for a drug can be framed as a LP task [5] but it is crucial that stakeholders comprehend the predictions before relying on them for making decisions about funding of research on the drug.

LP eXplanation (LP-X) methods [6] address this issue. Specifically, a post-hoc (after the prediction) LP-X method, works with a generic LP method and explains a prediction by selecting pieces of knowledge

XAI-KRKG@ECAI25: First International ECAI Workshop on eXplainable AI, Knowledge Representation and Knowledge Graphs, October 25–30, 2025, Bologna, Italy

*Corresponding author.
†These authors contributed equally.
r.barile17@phd.uniba.it (R. Barile); claudia.damato@uniba.it (C. d'Amato); l.santovito3@studenti.uniba.it (L. Santovito); nicola.fanizzi@uniba.it (N. Fanizzi)
https://rbarile17.github.io/robertobarile.github.io/ (R. Barile)
0009-0007-3058-8692 (R. Barile); 0000-0001-7116-9338 (C. d'Amato); 0000-0002-9421-8566 (N. Fanizzi)

published 2026-02-24

(e.g., sets of facts) that are associated to the prediction.

Nevertheless, multiple protocols exists for evaluating explanations, making it difficult the comparison of solutions coming from different LP-X methods. The prominent protocols for evaluating LP-X methods are re-training, introduced for evaluating CRIAGE [7] and also used for evaluating KELPIE [8] and KELPIE++ [9], and recall and support, proposed for evaluating CROSSE [10] and also used for evaluating SEMANTICCROSSE [11]. Conducting such evaluations is challenging because there is not yet consensus on a standard evaluation protocol for evaluating explanations and there are no insights on the consistency/reliability of the inter-rater (inter-protocol) evaluation, i.e., the evaluation of the same LP-X method via different protocols. For this purpose, we investigate the following Research Question (**RQ**):

Are the evaluation protocols re-training, recall, and support consistent when evaluating the same LP-X method?

Measuring the consistency between such evaluation protocols requires comparing the values they produce when applied to the same LP-X method. However, no state-of-the-art LP-X method has been evaluated with all of the prominent protocols. This is because the lack of a standard evaluation protocol has led to a proliferation of different ones, sometimes also tailored to the specific LP-X method to be evaluated. Specifically, in this paper, we address the RQ via evaluating the LP-X methods CROSSE and SEMANTICCROSSE, that can be readily evaluated via recall and support, also via re-training. This results in an evaluation of the same methods (CROSSE and SEMANTICCROSSE) under three different protocols, thus allowing to answer our RQ particularly by computing the correlation between the metrics resulting from the different protocols.

The rest of the paper is organized as follows. In § 2, we review state-of-the-art methods for computing and evaluating explanations. In § 3, we illustrate essential basics. In § 4, we detail the method for measuring the consistency between the evaluation protocols, while in § 5, we illustrate the experiments. In § 6, we summarize the achievements and suggest future research.

2. Related Works

This section analyzes the state-of-the-art LP-X methods along with the methods or protocols used for evaluating their performance. The LP-X methods, that we target because they are generic with respect to the LP method, explain a prediction by computing pieces of knowledge (e.g., sets of facts) that are associated to the prediction. The first proposals explain a prediction by returning exactly one fact (within the KG), as in the case of DP [12], applying perturbations, or CRIAGE [7] computing (approximate) influence functions. The latter can be restricted to a limited set of facts and to specific classes of KGE models. More recent methods explain a prediction by returning a set of facts. KELPIE [8] and KELPIE++ [9] employ a *post-training* process. KE-X [13] is based on information gain and KGEXPLAINER [14] adopts greedy search and perturbations. Notably, GENI [15] returns explanations also including ontological axioms based on numerical criteria on (specific classes of) KGEs. Conversely, the method introduced in [16] grounds on abduction via learned rules. The resulting explanations are mainly evaluated by *re-training* the KGE model, i.e., by comparing the LP performance of the original model with that of a model trained on a modified KG where the facts in the explanations have been added, removed, or isolated.

CROSSE [10] and SEMANTICCROSSE [11] explain a prediction by identifying a path between the entities in the prediction. They rely on similarity measures and evaluate explanations as the number of similar paths connecting similar entities. Other methods return explanations other than sets of facts or paths. For example, in [17] logical rules are mined to explain a set of predictions and are evaluated in terms of classification performance on the explained predictions and synthetic negative (false) facts. With FEABI [18], interpretable vectors are extracted from KGEs via feature selection and are compared to those learned with an interpretable LP method. The evaluation measures the influence of the LP explanations

on the solution of related tasks, without considering the user's perspective. These evaluation protocols do not allow comparing the explanations coming from the different approaches.

Another direction for evaluating explanations is to provide datasets containing ground-truth explanations to be compared with the computed ones. FR200K [19], FRUNI and FTREE [20], include hand-crafted rules that reflect domain knowledge and explain a fact by identifying those facts that underpin the rules generating it. In FR200K each explanation is also rated by users in terms of (subjective) intuitiveness, whereas in FRUNI and FTREE explanations are assumed to be valuable. Hence, FR200K enables user guided evaluation; however, its construction process hardly generalizes to large scale due to the required manual intervention.

A complementary direction is represented by interpretable LP methods, which are LP methods with a more understandable functioning. A comparison of different interpretable methods would mean to compare their functioning, and is beyond our purpose.

3. Basics

A KG $\mathcal{G}(\mathcal{V}, \mathcal{R})$ is a graph-based data structure, where $\mathcal{V}$ is a set of nodes representing entities, and $\mathcal{R}$ is a set of predicates, representing binary relations between entities. A KG can be seen as a collection of triples $\langle s, p, o\rangle \in \mathcal{V} \times \mathcal{R} \times \mathcal{V}$, with a *subject* s, a *predicate* p and an *object* o, where $s, o \in \mathcal{V}$ and $p \in \mathcal{R}$.

LP methods calculate a ranking function $\mathrm{rank}\colon \mathcal{V} \times \mathcal{R} \times \mathcal{V} \to \mathbb{N}$ that computes the position of a given triple $\langle s, p, o\rangle$ in the set of triples $\{\, \langle s, p, u\rangle \mid u \in \mathcal{V} \,\}$ according to the confidence/plausibility score computed via a KGE model. A triple in the KG $\mathcal{G}$ is correctly predicted by the LP method if it is top-ranked. The LP performance is typically evaluated in terms of the metrics:

- MRR: the average of the inverse of the obtained ranks
- $H@1$: the ratio of predictions for which the rank is 1

Next, let $explain\colon \mathcal{G} \to \mathcal{X}$ be the function denoting a LP-X method, where $\mathcal{X}$ is the set of all possible explanations. For example, the LP-X methods CrossE and SemanticCrossE that we specifically target, compute explanations as paths (maximum length 2) connecting the entities in the prediction. There are 6 possible types of path for a prediction $\langle s, p, o\rangle \in \mathcal{G}$:

1. $\{\, \langle s, p', o\rangle \,\}$;
2. $\{\, \langle o, p', s\rangle \,\}$;
3. $\{\, \langle u, p', s\rangle, \langle u, r, o\rangle \,\}$;
4. $\{\, \langle u, p', s\rangle, \langle o, r, u\rangle \,\}$;
5. $\{\, \langle s, p', u\rangle, \langle u, r, o\rangle \,\}$;
6. $\{\, \langle s, p', u\rangle, \langle o, r, u\rangle \,\}$.

where p' is a predicate similar to p, r is any other predicate $r \in \mathcal{R}$, and u is any other entity $u \in \mathcal{V}$. CrossE and SemanticCrossE return the empty set $\emptyset$ when they fail to explain the prediction $\langle s, p, o\rangle$. For computing similar predicates p', CrossE adopts the euclidean distance, whereas SemanticCrossE can adopt either the cosine distance or a semantic similarity measure.

4. The Proposed Approach

In this section, we illustrate the evaluation protocols to be compared, namely: re-training (§ 4.1), recall and support (§ 4.2), and how we verify the consistency, if any, between them in order to answer **RQ** (§ 4.3). We specifically consider and compare explanations only of the $T \subset \mathcal{G}$ of correct predictions made via a KGE model m, since explanations for wrong predictions may be misleading.

4.1. Re-training

The re-training protocol (introduced for evaluating the LP-X method CRIAGE [7]) measures the importance of the explanations by measuring the impact of explanations removal from the KG in solving the very same LP task via a KGE model m. Specifically, it is based on comparing the LP performance of the model used for computing the predictions with that of the model trained on a modified KG from which the triples in the explanations have been removed. If the removal significantly worsens performance, it indicates that the explanations are important for the predictions and as such they could be considered in principle as valid explanations.

In the following, we formalize[1] the re-training process, by considering a KG $\mathcal{G}(\mathcal{V}, \mathcal{R})$. First, let $\text{remove}_{\mathcal{G}} \colon 2^{\mathcal{X}} \to \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ be the function that removes the triples in each explanation x in a set of explanations X from the KG, formally:

$$\forall X \in 2^{\mathcal{X}}, \mathcal{G}' := \text{remove}_{\mathcal{G}}(X) = \mathcal{G} \setminus \bigcup_{x \in X} x$$

Second, let m' denote the perturbed KGE model, with the same architecture and hyperparameters as the KGE model m, but trained on the modified KG $\mathcal{G}'$ (instead of $\mathcal{G}$). $MRR_{m'}$ and $H@1_{m'}$ denote the LP performance metrics of the perturbed KGE model m'. Since the LP performance metrics MRR_m and $H@1_m$ of the original KGE model m are both 1.0 (since only the correct predictions are considered), the re-training metrics ΔMRR and $\Delta H@1$ can be computed as follows:

$$\Delta MRR = 1 - MRR_{m'}$$

$$\Delta H@1 = 1 - H@1_{m'}$$

Both fall within the interval $[0, 1]$, where higher values indicate more effective explanations.

4.2. Recall and Support

The recall (introduced for evaluating the LP-X method CROSSE) is the ratio of predictions for which the LP-X method generated an explanation, formally:

$$\forall T \subset \mathcal{G}, \text{recall}(T) = \frac{|\{\, t \mid t \in T, \text{explain}(t) \neq \emptyset \,\}|}{|T|}$$

The support of an explanation for a prediction, introduced for evaluating the LP-X method CROSSE, is the number of supports, i.e., triples in the KG that are similar to the prediction and have an explanation similar to the one of the prediction. The support is computed whilst computing the explanation and is returned along with the explanation because CROSSE and SEMANTICCROSSE returns solely the explanations with at least one support. Moreover, different similarity functions are defined in CROSSE (euclidean) and SEMANTICCROSSE (cosine and semantic). In the following, we formalize the support, considering a KG $\mathcal{G}(\mathcal{V}, \mathcal{R})$. First, let $\text{neighbors}_k \colon \mathcal{V} \to \mathcal{V}^k$ be the function selecting the k entities that are most similar to the given entity. Second, let $\text{get\_sim\_triples}_k \colon \mathcal{G} \to 2^{\mathcal{G}}$ be the function that selects the triples similar to the given one, via the function neighbors_k: $\forall \langle s, p, o \rangle \in \mathcal{G}$

$$\text{sim\_triples}_k(\langle s, p, o \rangle) = \{\, \langle n, p, v \rangle \mid n \in \text{neighbors}_k(s) \wedge v \in \mathcal{V} \wedge \langle n, p, v \rangle \in \mathcal{G} \,\}.$$

Next, let $\text{is\_support} \colon \mathcal{G} \times \mathcal{X} \times \mathcal{G} \times \mathcal{X} \to \{\, 0, 1 \,\}$ be the function that determines whether an explanation for a given prediction is supported by another explanation for another given prediction. Formally, given a prediction $\langle s, p, o \rangle$ with its corresponding explanation x_1 and a similar triple $\langle n, p, v \rangle$ with its corresponding explanation x_2, we specify for each possible type of the explanation x_1, when x_2 is a support:

1. $x_1 = \{\, \langle s, p', o \rangle \,\}, x_2 = \{\, \langle n, p', v \rangle \,\}$;

[1] We denote the set of all the subsets of a set A as 2^A.

2. $x_1 = \{ \langle o, p', s \rangle \}, x_2 = \{ \langle v, p', n \rangle \}$;
3. $x_1 = \{ \langle u, p', s \rangle, \langle u, r, o \rangle \}, x_2 = \{ \langle u, p', n \rangle, \langle u, r, v \rangle \}$;
4. $x_1 = \{ \langle u, p', s \rangle, \langle o, r, u \rangle \}, x_2 = \{ \langle u, p', n \rangle, \langle v, r, u \rangle \}$;
5. $x_1 = \{ \langle s, p', u \rangle, \langle u, r, o \rangle \}, x_2 = \{ \langle n, p', u \rangle, \langle u, r, v \rangle \}$;
6. $x_1 = \{ \langle s, p', u \rangle, \langle o, r, u \rangle \}, x_2 = \{ \langle n, p', u \rangle, \langle v, r, u \rangle \}$.

where p' is a predicate similar to p, r is any other predicate $r \in \mathcal{R}$, and u is any other entity $u \in \mathcal{V}$.

Then, let $\text{support}_k \colon \mathcal{G} \times \mathcal{X} \to \mathbb{N}$ be the function that measures the number of supports for the explanation of a given prediction, formally: $\forall \langle s, p, o \rangle \in \mathcal{G}, x = \text{explain}(\langle s, p, o \rangle)$

$$\text{support}_k(\langle s, p, o \rangle, x) = |\{ \langle n, p, e \rangle \in \text{sim\_triples}(\langle \text{s, p, o} \rangle) \wedge \text{is\_support}(\langle s, p, o \rangle, x, \langle n, p, e \rangle, \text{explain}(\langle n, p, e \rangle) \}|$$

Finally, let $T = [t_1, ..., t_n] \subset \mathcal{G}$ be a sequence of predicted triples and $X = [x_1, \ldots, x_n] \subset \mathcal{X}$ be a sequence of explanations such that $\forall i \in \{ 1, \ldots, n \}$ explanation x_i explains prediction t_i ($\text{explain}(t_i) = x_i$), then let $\text{average\_support} \colon 2^{\mathcal{X}} \times 2^{\mathcal{G}} \to \mathbb{R}$ be the function measuring the average support of a sequence of predictions with their corresponding explanations, formally:

$$\text{average\_support}(T, X) = \frac{1}{|X|} \sum_{i=1}^{n} \text{support}(t_i, x_i)$$

Higher values of average support indicate more effective explanations. Specifically, the sequences of explanations and predictions are partitioned into six disjoint subsets, one for each explanation type. The average support is then computed independently for each subset.

4.3. Measuring the Consistency between the Evaluation Protocols

To verify the consistency between the protocols, we employ the standard Pearson correlation coefficient, as it has been used to assess the inter-rater consistency among raters/evaluators using continuous scales.

The Pearson correlation coefficient ρ measures the linear relationship between two sets of variables A and B and falls within the interval $[-1, 1]$, where 1 indicates a perfect positive correlation (as A increases, so does B), -1 indicates perfect negative correlation (as A increases, B decreases), and 0 indicates that there is no linear relationship between the variables.

As for the support protocol, in addition to the average_support for each explanation type, we compute the total number of supports ($\#\text{supports}$) for the set of evaluated explanations, since we consider the quality of an explanation to be independent of the type of path, and dependent solely on the number of supports. Hence, we compute the correlation between the following pairs of metrics:

- ΔMRR and recall;
- $\Delta H@1$ and recall;
- ΔMRR and $\#\text{supports}$;
- $\Delta H@1$ and $\#\text{supports}$.

For each correlation value, we also perform a permutation test that outputs a p-value intuitively denoting the probability that the correlation is due to chance: $p < 0.05$ denotes a statistically significant (not due to chance) correlation.

5. Experimental Evaluation

In this section, we illustrate the experimental setup (§ 5.1) and discuss the results (§ 5.2).

Table 1
KG statistics

	Entities	Predicates	Train triples	Valid triples	Test triples
DB100K	98776	464	587688	49172	49114
YAGO4-20	96910	70	555182	69398	69398

5.1. Experimental Setup

We performed the study on two publicly available KGs: YAGO4-20, DB100K sampled from DBpedia and YAGO4, respectively [9]; their statistics are reported in Tab. 1. YAGO4-20 and DB100K contain not only triples, but also ontological axioms that SemanticCrossE leverages for computing the explanations via the semantic similarity measure. In addition, we performed the experiments with respect to three different LP methods. Specifically, we adopted three seminal KGE models, each representing a prominent family of such models, namely: TransE [21] (translational), ConvE [22] (neural) and ComplEx [23] (tensor factorization). Moreover, since KGE models are machine learning solutions, the KGs are further split into a training set, a validation set, and a test set of triples. For each KGE model and KG, we computed the explanations via CrossE and SemanticCrossE of the triples in the test set that are (correctly) top-ranked via the LP method. We employ solely CrossE and SemanticCrossE as it is difficult to evaluate the other SOTA methods, such as Criage and Kelpie, via the recall and support protocols. The number of explained triples for each KGE model and KG is reported in Tab. 2. All the code, datasets, and trained models utilized in our study are openly accessible on GitHub[2]. The correlations are computed firstly for the complete set of results and then considering separately the results for each KGE model and each KG.

Table 2
Number of correct predictions

KGE model	KG	# Correct Predictions
TransE	DB100K	8277
TransE	YAGO4-20	9288
ConvE	DB100K	18848
ConvE	YAGO4-20	16655
ComplEx	DB100K	19276
ComplEx	YAGO4-20	18399

5.2. The Outcomes of the Evaluation

Tab. 3 reports the outcomes of the evaluation of the computed explanations via the protocols re-training, recall, and support. Based on such values, we computed the correlation coefficients, reported in Tab. 4. The correlation coefficients suggest a moderate and significant positive correlation for all pairs of metrics. As for the analysis conducted separately for each KGE model, the outcomes suggest a strong and significant positive correlation when considering ConvE and ComplEx, but not when considering TransE. Specifically, considering TransE, the correlation of the re-training metrics with the recall is close to 0 and not significant, while the one of the re-training metrics with the number of supports is strongly negative and significant. The low correlation when considering TransE may be due to the lower performance, in terms of the number of correct predictions in Tab. 2, of such a model compared to that of the other models. As for the analysis conducted separately for each KG, the outcomes suggest a strong and significant positive correlation when considering DB100K and a correlation close to 0 and not significant when considering YAGO4-20. The low correlation when considering YAGO4-20 may be

[2]https://github.com/LeoSantovito/lpx_evalprotocol_consistency

Table 3
Evaluation of the explanations via the protocols re-training, recall, and support

KGE model	KG	LP-X method	Similarity	ΔMRR	$\Delta H@1$	recall	#supports
TransE	DB100K	CrossE	euclidean	0.386	0.522	0.077	12710
TransE	DB100K	SemanticCrossE	cosine	0.363	0.494	0.094	12022
TransE	DB100K	SemanticCrossE	semantic	0.360	0.483	0.098	14706
ConvE	DB100K	CrossE	euclidean	0.746	0.828	0.077	123876
ConvE	DB100K	SemanticCrossE	cosine	0.770	0.844	0.124	159068
ConvE	DB100K	SemanticCrossE	semantic	0.764	0.842	0.147	152723
ComplEx	DB100K	CrossE	euclidean	0.857	0.910	0.241	320007
ComplEx	DB100K	SemanticCrossE	cosine	0.843	0.898	0.326	492067
ComplEx	DB100K	SemanticCrossE	semantic	0.825	0.993	0.305	434052
TransE	YAGO4-20	CrossE	euclidean	0.489	0.576	0.090	5061
TransE	YAGO4-20	SemanticCrossE	cosine	0.479	0.571	0.089	7525
TransE	YAGO4-20	SemanticCrossE	semantic	0.492	0.579	0.104	6761
ConvE	YAGO4-20	CrossE	euclidean	0.565	0.647	0.001	53
ConvE	YAGO4-20	SemanticCrossE	cosine	0.575	0.657	0.042	3473
ConvE	YAGO4-20	SemanticCrossE	semantic	0.642	0.703	0.056	3576
ComplEx	YAGO4-20	CrossE	euclidean	0.788	0.829	0.061	5761
ComplEx	YAGO4-20	SemanticCrossE	cosine	0.785	0.827	0.074	6898
ComplEx	YAGO4-20	SemanticCrossE	semantic	0.794	0.834	0.078	6589

Table 4
The correlations of the different pairs of metrics. Significant correlations ($p < 0.05$) are marked with *

Set	Metric	ΔMRR - recall	$\Delta H@1$ - recall	ΔMRR - #supports	$\Delta H@1$ - #supports
Complete	ρ	0.491*	0.522*	0.612*	0.649*
	p-value	0.038	0.026	0.007	0.004
TransE	ρ	0.171	0.042	-0.964*	-0.945*
	p-value	0.746	0.937	0.002	0.004
ConvE	ρ	**0.909***	**0.901***	**0.961***	**0.978***
	p-value	0.012	0.014	0.002	0.000
ComplEx	ρ	**0.865***	**0.912***	**0.864***	**0.911***
	p-value	0.026	0.011	0.026	0.011
DB100K	ρ	**0.686***	**0.668***	**0.818***	**0.805***
	p-value	0.041	0.049	0.007	0.009
YAGO4-20	ρ	-0.127	-0.147	0.170	0.157
	p-value	0.744	0.706	0.660	0.687

due to the low number of supports (compared to DB100K) that in turn may be due to the low number of predicates in YAGO4-20.

6. Conclusions

We conducted an empirical study of the consistency between the prominent protocols for evaluating explanation, namely re-training, recall, and support. Specifically, we evaluated CrossE and SemanticCrossE, originally evaluated via recall and support, also via re-training. Hence, we computed the correlation between the metrics resulting from the different protocols. The outcomes suggest that the protocols are overall consistent. A current limitation stems from the number of explained predictions

that varies across KGE models and KGs. For the future, we aim not only at conducting a study with a fixed number of explained predictions, but also at extending the study with other protocols, such as LP-DIXIT [24], other KGs, including those without schema level knowledge, and other consistency statistics.

Acknowledgments

This work was partially supported by project *FAIR - Future AI Research* (PE00000013), spoke 6 - Symbiotic AI (https://future-ai-research.it/) under the PNRR MUR program funded by the European Union - NextGenerationEU, and by PRIN project *HypeKG - Hybrid Prediction and Explanation with Knowledge Graphs* (Prot. 2022Y34XNM, CUP H53D23003700006) under the PNRR MUR program funded by the European Union - NextGenerationEU

Declaration on Generative AI

The authors have not employed any Generative AI tool.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, ACM Computing Surveys 54 (2022) 1–37. doi:10.1145/3447772.

[2] T. Pellissier Tanon, G. Weikum, F. Suchanek, YAGO 4: A Reason-able Knowledge Base, in: A. Harth, et al. (Eds.), The Semantic Web, volume 12123, Berlin, Heidelberg, 2020, pp. 583–596. doi:10.1007/978-3-030-49461-2_34.

[3] X. L. Dong, Building a broad knowledge graph for products, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE Computer Society, Washington DC, USA, 2019, pp. 25–25. doi:10.1109/ICDE.2019.00010.

[4] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo, Knowledge Graph Embedding for Link Prediction: A Comparative Analysis, ACM Transactions on Knowledge Discovery from Data 15 (2021) 1–49. doi:10.1145/3424672.

[5] V. Nováček, S. K. Mohamed, Predicting Polypharmacy Side-Effects using Knowledge Graph Embeddings, AMIA Summits on Translational Science Proceedings 2020 (2020) 449.

[6] S. Schramm, C. Wehner, U. Schmid, Comprehensible Artificial Intelligence on Knowledge Graphs: A survey, Journal of Web Semantics 79 (2023) 100806.

[7] P. Pezeshkpour, C. A. Irvine, Y. Tian, S. Singh, Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications, in: Burstein, Jill, Doran, Christy, Solorio, Thamar (Eds.), NAACL-HLT '19: 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); Minneapolis, Minnesota, USA; 02-07 June 2019, volume 1, Association for Computational Linguistics, Kerrville, Texas, USA, 2019, pp. 3336–3347. doi:10.18653/v1/N19-1337.

[8] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining Link Prediction Systems based on Knowledge Graph Embeddings, in: Z. Ives (Ed.), SIGMOD/PODS '22: Proceedings of the 2022 International Conference on Management of Data; Philadelphia, Pennsylvania, USA; 12-17 June 2022, ACM, New York, New York, USA, 2022, pp. 2062–2075. doi:10.1145/3514221.3517887.

[9] R. Barile, C. d'Amato, N. Fanizzi, Explanation of Link Predictions on Knowledge Graphs via Levelwise Filtering and Graph Summarization, in: A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), Proceedings of the 26th European

Semantic Web Conference (ESWC 2024), volume 14664, Springer Nature Switzerland, Cham, 2024, pp. 180–198. doi:10.1007/978-3-031-60626-7_10.

[10] W. Zhang, B. Paudel, W. Zhang, A. Bernstein, H. Chen, S. Culpepper, Interaction Embeddings for Prediction and Explanation in Knowledge Graphs, in: WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, February 11-15, 2019, ACM, New York, New York, USA, 2019, pp. 96–104. doi:10.1145/3289600.3291014.

[11] C. d'Amato, P. Masella, N. Fanizzi, An Approach Based on Semantic Similarity to Explaining Link Predictions on Knowledge Graphs, in: J. He, R. Unland, E. J. Santos, X. Tao, H. Purohit, W.-J. van den Heuvel, J. Yearwood, J. Cao (Eds.), IEEE/WIC/ACM International Conference on Web Intelligence, ACM, New York, New York, USA, 2021, pp. 170–177. doi:10.1145/3486622.349395.

[12] H. Zhang, T. Zheng, J. Gao, C. Miao, L. Su, Y. Li, K. Ren, Data Poisoning Attack against Knowledge Graph Embedding, in: S. Kraus (Ed.), IJCAI '19: Proceedings of the 28th International Joint Conference on Artificial Intelligence; Macao, China; 10-16 August 2019, IJCAI, Online, 2019, pp. 4853–4859. doi:10.24963/ijcai.2019/674.

[13] D. Zhao, G. Wan, Y. Zhan, Z. Wang, L. Ding, Z. Zheng, B. Du, KE-X: Towards subgraph explanations of knowledge graph embedding based on knowledge information gain, Knowledge-Based Systems 278 (2023) 110772. doi:10.1016/j.knosys.2023.110772.

[14] T. Ma, X. song, W. Tao, M. Li, J. Zhang, X. Pan, J. Lin, B. Song, x. Zeng, KGExplainer: Towards Exploring Connected Subgraph Explanations for Knowledge Graph Completion, 2024. arXiv:2404.03893.

[15] E. Amador-Domínguez, E. Serrano, D. Manrique, GEnI: A framework for the generation of explanations and insights of knowledge graph embedding predictions, Neurocomputing 521 (2023) 199–212. doi:10.1016/j.neucom.2022.12.010.

[16] P. Betz, C. Meilicke, H. Stuckenschmidt, Adversarial Explanations for Knowledge Graph Embeddings, in: L. De Raedt (Ed.), IJCAI '22: Proceedings of the 31th International Joint Conference on Artificial Intelligence; Vienna, Austria; 23-29 July 2022, IJCAI, Online, 2022, pp. 2820–2826. doi:10.24963/ijcai.2022/391.

[17] N. A. Krishnan, C. R. Rivero, A Model-Agnostic Method to Interpret Link Prediction Evaluation of Knowledge Graph Embeddings, in: I. Frommholz (Ed.), CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, United Kingdom, October 21-25, 2023, ACM, New York, New York, USA, 2023, pp. 1107–1116. doi:10.1145/3583780.3614763.

[18] Y. Ismaeil, D. Stepanova, T.-K. Tran, H. Blockeel, FeaBI: A Feature Selection-Based Framework for Interpreting KG Embeddings, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, Springer, Cham, Switzerland, 2023, pp. 599–617. doi:10.1007/978-3-031-47240-4_32.

[19] N. Halliwell, F. Gandon, F. Lecue, User Scored Evaluation of Non-Unique Explanations for Relational Graph Convolutional Network Link Prediction on Knowledge Graphs, in: Proceedings of the 11th Knowledge Capture Conference, ACM, Virtual Event USA, 2021, pp. 57–64. doi:10.1145/3460210.3493557.

[20] P. S. Martin, T. Besold, P. Kumari, FRUNI and FTREE Synthetic Knowledge Graphs for Evaluating Explainability, in: XAI in Action: Past, Present, and Future Applications@NeurIPS2023 (No Formal Proceedings), OpenReview, Online, 2023.

[21] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-Relational Data, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 2787–2795.

[22] T. Dettmers, P. Minervini, P. Stenetorp, S. Riedel, Convolutional 2d knowledge graph embeddings, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, Cambridge,

Massachusetts, 2018, pp. 1811–1818. URL: https://https://dl.acm.org/doi/10.5555/3504035.3504256.

[23] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: International Conference on Machine Learning, JMLR, Online, 2016, pp. 2071–2080. doi:10.5555/3045390.3045609.

[24] R. Barile, C. d'Amato, N. Fanizzi, LP-DIXIT: Evaluating Explanations for Link Predictions on Knowledge Graphs using Large Language Models, in: Proceedings of the ACM on Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 4034–4042. URL: https://doi.org/10.1145/3696410.3714667. doi:10.1145/3696410.3714667.