Effects of Locality and Rule Language on Explanations for Knowledge Graph Embeddings

Luis Galárraga

Effects of Locality and Rule Language on Explanations for Knowledge Graph Embeddings

Luis Galárraga[1][0000−0002−0241−5379]

Inria, France
luis.galarraga@inria.fr

Abstract. Knowledge graphs (KGs) are key tools in many AI-related tasks such as reasoning or question answering. This has, in turn, propelled research in link prediction in KGs, the task of predicting missing relationships from the available knowledge. Solutions based on KG embeddings have shown promising results in this matter. On the downside, these approaches are usually unable to explain their predictions. While some works have proposed to compute post-hoc rule explanations for embedding-based link predictors, these efforts have mostly resorted to rules with unbounded atoms, e.g., $bornIn(x, y) \Rightarrow residence(x, y)$, learned on a global scope, i.e., the entire KG. None of these works has considered the impact of rules with bounded atoms such as $nationality(x, England) \Rightarrow speaks(x, English)$, or the impact of learning from regions of the KG, i.e., local scopes. We therefore study the effects of these factors on the quality of rule-based explanations for embedding-based link predictors. Our results suggest that more specific rules and local scopes can improve the accuracy of the explanations. Moreover, these rules can provide further insights about the inner-workings of KG embeddings for link prediction.

Keywords: knowledge graph embeddings, explainable AI

1 Introduction

The continuous advances in information extraction on the Web have given rise to large repositories of machine-friendly statements modeled as knowledge graphs (KGs). These are collections of facts of the form $p(s, o)$ that describe real-world entities, e.g., $capital(Italy, Rome)$. In this formalism, the predicate p in a statement $p(s, o)$ can be seen as a directed labeled edge that connects the subject s to the object o. KGs allow computers to "understand" the real world, and find applications in multiple AI-related tasks such as entity-centric IR, reasoning, question answering, smart assistants, etc. Since KGs usually suffer from incompleteness, a central task in KGs is *link prediction*, where the goal is to infer new facts from the available knowledge. Link prediction constitutes a fundamental step towards proper knowledge graph completion.

Approaches for link prediction in KGs abound and fall mainly into two paradigms. On the one hand, *symbolic methods* [11, 12, 16, 19] mine explicit patterns on the graph, e.g., the rule $capital(x, y) \Rightarrow inCountry(y, x)$, and use

those patterns to infer new relationships between entities. On the other hand, approaches based on latent factors [4, 21, 22, 29, 31, 34] embed predicates p and entities s, o in a latent space driven by a score function that ranks true facts better than false ones. For example, TransE [4] learns d-dimensional embeddings (in bold) for predicates and entities such that $\boldsymbol{s} + \boldsymbol{p} \approx \boldsymbol{o}$, if $p(s, o)$ holds in reality. TransE's score function for facts is then $-\|\boldsymbol{s} + \boldsymbol{p} - \boldsymbol{o}\|_l$ $(l = \{1, 2\})$.

Embedding-based methods have exhibited promising performance for link prediction, however their main downside is that they operate as black boxes: one cannot obtain an explanation of the logic behind a predicted fact $p(s, o)$ from the latent representations of s, p, and o. This has therefore motivated some works on mining rule-based explanations for KG embeddings [7, 9, 26, 27]. Those explanations can help us, for instance, verify if the embeddings meet expected reasoning guarantees such as transitivity, i.e., $p(x, z) \wedge p(z, y) \Rightarrow p(x, y)$, or detect biases in the data. It is known that redundancy in the form of inverse predicates, e.g., $hyponym(feline, cat)$, $hypernym(cat, feline)$ in benchmark datasets, led to over-estimated accuracies for state-of-the-art embedding-based link predictors [3, 18]. Had a mechanism to understand that the embeddings mainly captured patterns such as $hyponym(x, y) \Rightarrow hypernym(y, x)$, this issue could have been detected in advance.

A limitation of existing explanations for KG embeddings is that they only capture global inference patterns. This is tantamount to mining explanations in the language of unbounded atoms, i.e., rules with no constants in the arguments such as $bornIn(x, z) \wedge officialLang(z, y) \Rightarrow speaks(x, y)$, that hold *globally*, that is, on the entire KG. However, such rules cannot express specific entity associations such as $nationality(x, USA) \Rightarrow speaks(x, English)$, presumably captured by link predictors. On those grounds, Section 4 addresses the following research question (**RQ1**): **what is the impact of specific rules in the quality of the explanations for embedding-based predictors?**. Moreover, and in line with existing works in interpretable AI [23, 24], we also study a second research question (**RQ2**): **how does learning explanation rules on specific regions of the KG, i.e., local explanations, impact the quality of the resulting rules?**. Before answering these questions, we discuss basic concepts and related work in Section 2, and explain how to compute rule-based explanations for link predictors in Section 3.

2 Preliminaries

2.1 Background Concepts

Knowledge Graphs. A knowledge graph $\mathcal{K} = (\mathcal{V}, \mathcal{E}, l_v, l_e)$ is a directed labeled graph with sets of vertices $\mathcal{V}$ and edges $\mathcal{E}$, where the injective functions $l_v : \mathcal{V} \to \mathcal{I}$, $l_e : \mathcal{E} \to \mathcal{P}$ assign labels to the vertices and edges. The sets $\mathcal{I}$ and $\mathcal{P}$ contain entity and predicate labels. An edge labeled *capital* departing from a vertex labeled *France* to a vertex labeled *Paris* denotes the *statement* or *fact* $capital(France, Paris)$. Hence, a KG $\mathcal{K} \subset \mathcal{I} \times \mathcal{P} \times \mathcal{I}$ is also a set of facts $p(s, o)$

with subject s, predicate p, and object o. Usually, standard KGs store only facts believed to be true.

We define the *potential set* $\Omega(\mathcal{K})$ of a KG as the universe of facts that could be constructed from the entities and predicates in $\mathcal{K}$. More formally, $\Omega(\mathcal{K}) = \mathcal{D}_v(\mathcal{K}) \times \mathcal{D}_e(\mathcal{K}) \times \mathcal{D}_v(\mathcal{K})$ where

$$\mathcal{D}_v(\mathcal{K}) = \{l_v(v) : v \in \mathcal{V}\}, \qquad \mathcal{D}_e(\mathcal{K}) = \{l_e(e) : e \in \mathcal{E}\}$$

are the *entity and predicate domains* of $\mathcal{K}$. Furthermore, we define the *potential set* of a predicate p as $\Omega(\mathcal{K}) \supseteq \Omega^p(\mathcal{K}) = \{p(s,o) : (s,o) \in \mathcal{D}^p(\mathcal{K}) \times \bar{\mathcal{D}}^p(\mathcal{K})\}$ with

$$\mathcal{D}^p(\mathcal{K}) = \{s : \exists o : p(s,o) \in \mathcal{K}\}, \qquad \bar{\mathcal{D}}^p(\mathcal{K}) = \{o : \exists s : p(s,o) \in \mathcal{K}\}.$$

$\Omega^p(\mathcal{K})$ therefore defines the set of all possible facts that could be constructed with the known subjects and objects of predicate p.

Horn rules. An *atom* A is a statement with constant predicate such that its subject and object arguments can be variables $v \in \mathbb{V}$ with $\mathbb{V} \cap \mathcal{I} = \emptyset$. If A has only variable arguments, we say A is *unbounded*, otherwise it is *bounded*. A *Horn rule* R is a statement of the form $\boldsymbol{B} \Rightarrow H$ where the *body* $\boldsymbol{B}$ is a conjunction of atoms $\bigwedge_{1 \leq i \leq n} A_i$, and H is the head atom. For instance, the rule $\mathit{parent}(x,z) \wedge \mathit{nationality}(z,y) \Rightarrow \mathit{nationality}(x,y)$ states that parents and children have the same nationality. These rules usually come with scores that quantify their precision. It is common to require atoms in rules to have at least one variable, be transitively connected, and form *safe* rules, that is, ensure that the head variables occur also in the body. This condition guarantees that the head variables are universally quantified, allowing for concrete predictions via *substitutions*. A substitution $\sigma : \mathbb{V} \to \mathcal{I}$ is a partial mapping from variables to constants, such that its application to atoms or rules replaces each variable with its corresponding constant in the mapping. For example, applying the substitution $\sigma = \{x \to \mathit{Marie\ Curie}, y \to \mathit{France}\}$ to the atom $A : \mathit{nationality}(x,y)$, gives a new atom $\sigma(A) : \mathit{nationality}(\mathit{Marie\ Curie}, \mathit{France})$. We say a rule $R : \boldsymbol{B} \Rightarrow H$ *predicts a fact* A' in a KG $\mathcal{K}$, denoted by $R \wedge \mathcal{K} \vdash A'$, iff $\exists\, \sigma : (\forall B \in \boldsymbol{B} : \sigma(B) \in \mathcal{K}) \wedge \sigma(H) = A'$. Put differently a rule predicts a fact A' if there exist a substitution σ that (i) maps each atom in the rule's body to a known KG fact, and (ii) maps the head atom to A'. If R predicts a statement A' and $A' \in \mathcal{K}$, we say that R *predicts* A' *correctly*, i.e., the prediction is a known fact, and we use the notation $R \wedge \mathcal{K} \vDash A'$.

Link Predictors. A *link predictor* $f : \Omega(\mathcal{K}) \to \mathbb{R}$ is a function that scores the facts in the potential set of a KG, usually assigning higher values to true facts. Link predictors are mostly used to answer queries of the forms $p(s, ?)$ or $p(?, o)$, in other words, queries that ask for the most likely subject or object of a statement given the other two components. Embedding-based link predictors operate on latent representations for entities, predicates, and facts in $\Omega(\mathcal{K})$.

Hence, they actually have the form $f = \hat{f} \circ h$, where $\hat{f} : \mathbb{C}^k \to \mathbb{R}^1$ is a function defined on a k-dimensional representation for facts, and $h : \Omega(\mathcal{K}) \to \mathbb{C}^k$ maps facts to k-dimensional vectors. If the semantics of the vector components are not understandable to humans, we say that f is a black box. That is the case for pure embedding-based link predictors such as TransE [4] or ComplEx [31].

Explanations. An explanation $E = \langle \mathcal{R}, g \rangle$ for a black-box link predictor $f : \Omega(\mathcal{K}) \to \mathbb{R}$ consists of a set $\mathcal{R}$ of Horn rules and a function $g : \mathcal{R} \to \mathbb{R}$ that attributes higher scores to rules that "agree" with f. A rule $R : \boldsymbol{B} \Rightarrow H$ agrees with f, if R predicts a fact $A \in \Omega(\mathcal{K})$ also predicted by f. This definition assumes the existence of a threshold θ such that $f(A) \geq \theta$ is interpreted as the black box also "thinking" that A is true. Explanations can be of different scope, namely *global* when they are learned on the potential set $\Omega^p(\mathcal{K})$ of a predicate p, or *local* when they are learned on smaller regions of $\Omega^p(\mathcal{K})$ as explained in Section 3.

2.2 Related Work

Link Prediction. This problem has received a lot of attention in the last 10 years with approaches lying on a spectrum from symbolic methods to embedding-based techniques. We refer the reader to [14] for a comprehensive survey. Symbolic techniques learn explicit patterns, e.g., arbitrary subgraphs, paths, association rules, Horn rules, etc., from KGs and use those patterns as features to predict missing links between entities [11, 12, 16, 19]. In contrast, the common principle of embedding-based methods is to model entities and predicates as elements in a latent space, where predicates characterize interactions between entity embeddings. Those interactions are modeled as geometrical operations, e.g., translation in TransE [4] where $\boldsymbol{s} + \boldsymbol{p} \approx \boldsymbol{o}$ for true facts $p(s, o)$ $(\boldsymbol{s}, \boldsymbol{p}, \boldsymbol{o} \in \mathbb{R}^d)$, or rotation in RotatE [30]. More recent methods resort to neural architectures [20, 28] that exploit the vicinity of entities in the graph to learn proper latent representations for both entities and predicates.

In all cases, a scoring function – implemented by minimizing a loss function – guides the training of the embeddings, which are learned to yield high scores for true facts and low scores for false facts. The latter are obtained by corrupting the true facts in the KG – a task of utter importance for the quality of the embeddings [13, 36].

Other methods combine the strengths of symbolic patterns and embeddings [6]. In [17], the authors improve the accuracy of different state-of-the-art embedding-based link predictors by removing those predictions that are not backed up by any of the Horn rules learned on the data. This strategy is complemented with a combined ranking that takes into account the individual rankings given by the rules and the embeddings. Some approaches [1, 13, 35] propose iterative algorithms that use rules and embeddings to produce better examples for subsequent

[1] Most methods embed the entities in real spaces, i.e., in $\mathbb{R}^k$, but a few, e.g., [31] resort to vectors of complex numbers in $\mathbb{C}^k$.

training. In contrast, other methods [10, 25] instruct the embeddings to comply to explicit reasoning patterns, e.g., transitivity, $p(x, z) \wedge p(z, y) \Rightarrow p(x, y)$.

Explaining the Black Box. Unlike symbolic approaches, link predictors based on embeddings are black boxes. Hence, there have been some efforts to explain their logic by mining explicit patterns [7, 26, 27] with attribution scores. Among these patterns, Horn rules are the most expressive. The rules are extracted using state-of-the-art rule or path mining algorithms [2,8,15,16], whereas the attribution scores are learned via machine learning, e.g., linear or logistic regression in the spirit of standard explanation techniques such as LIME [24]. Nevertheless, none of these approaches exploits the power of Horn rules at its best. For instance, [7, 27] mine rule explanations of up to two atoms, e.g., $bornIn(x, y) \Rightarrow livesIn(x, y)$, whereas DistMult [34] can only learn *pure* paths such as $bornIn(x, z) \wedge inCountry(z, y) \Rightarrow nationality(x, y)$. Hence, none of these methods can induce explanations in the language of bounded atoms such as $nationality(x, UK) \Rightarrow speaks(x, English)$. Furthermore, all these endeavors mine global explanations. Embedding-based models can, though, be very complex and therefore hard to approximate in the general sense. Thus we explore the effects of bounded atoms and locality in the quality of the explanations.

3 Explaining KG Embeddings for Link Prediction

Algorithm 1 describes a generic procedure to compute rule-based explanations for a black-box link predictor f trained on a KG $\mathcal{K}$, containing both true ($\mathcal{K}^+$) and corrupted facts ($\mathcal{K}^-$), in line with existing approaches [7, 26, 27]. The rules are learned on a context C consisting of facts of a given predicate p. We elaborate on the stages of the algorithm and the different ways to define the context C.

Algorithm 1: Build Explanation

Input: link predictor $f : \Omega(\mathcal{K}) \rightarrow \mathbb{R}$ trained on $\mathcal{K} = \mathcal{K}^+ \cup \mathcal{K}^-$, context $C \subset \Omega^p(\mathcal{K})$, $C \cap \mathcal{K} = \emptyset$

Output: an explanation $E = \langle \mathcal{R}, g \rangle$ with set of rules $\mathcal{R}$, $g : \mathcal{R} \rightarrow \mathbb{R}$

```
1 K̂ := ∅
2 foreach A := p(s, o) ∈ C do
3     if f(A) ≥ θ then
4         K̂ := K̂ ∪ {p^f(s, o)}
5     else
6         K̂ := K̂ ∪ {¬p^f(s, o)}
7 R := rule mining on K ∪ K̂ for predicates p^f, ¬p^f
8 return build-rule-based-surrogate(R, K, K̂, f)
```

Binarizing the black box. To learn Horn rules that mimic a black box f, we need to convert f's scores for facts into true or false verdicts. To this end, lines 2–6 label each fact in the context C by computing f's score and then applying a threshold to decide whether the fact is deemed true or not by f. This set $\hat{\mathcal{K}}$ of annotated facts is represented by the surrogate predicates p^f, $\neg p^f$.

Rule Mining. Line 7 in Alg. 1 learns Horn rules of the forms $\boldsymbol{B} \Rightarrow p^f(s,o)$ and $\boldsymbol{B} \Rightarrow \neg p^f(s,o)$ with confidence scores from the original KG $\mathcal{K}$ and the black-box annotated context $\hat{\mathcal{K}}$.

Learning the explanation. Finally, line 8 uses the rules as features to learn a surrogate model $f_s : \mathbb{R}^{|\mathcal{R}|} \to \mathbb{R}$ that mimics the binarized f and provides importance scores for the rules in $\mathcal{R}$. Given a statement $A = \bar{p}^f(s,o) \in \hat{\mathcal{K}}$ with $\bar{p}^f \in \{p^f, \neg p^f\}$, we encode A as a vector $x_A \in \mathbb{R}^{|\mathcal{R}|}$ such that its i-th entry is set as follows:

$$x_A[i] = \begin{cases} sgn(A) \times conf(R_i) & R_i \wedge (\mathcal{K} \cup \hat{\mathcal{K}}) \vDash A \\ -sgn(A) \times conf(R_i) & R_i \wedge (\mathcal{K} \cup \hat{\mathcal{K}}) \vdash A' \text{ with } A' \neq A \\ 0 & \text{otherwise} \end{cases}$$

Here $sgn(A) = 1$ if $A = \bar{p}^f(s,o)$, otherwise $sgn(A) = -1$. If a rule $R_i \in \mathcal{R}$ *predicts correctly* a statement $A \in \hat{\mathcal{K}}$, the i-th component of x_A holds a value equals the confidence of R_i (reported by the rule mining phase) with the same polarity of f's prediction. In that case, R_i agrees with f and is a potential explanation for f's answer on A. If R_i is a potential explanation for some other fact A', we change the sign of confidence value. In any other case, we assign a score of 0 to the entry. We use the x_A vectors and the binarized labels – given by $sgn(A)$ – to train a surrogate logistic regression classifier f_s, whose coefficients define an attribution mapping $g : \mathcal{R} \to \mathbb{R}$ for rules – our explanation. The surrogate f_s can provide both binary labels and probability scores for facts, and its coefficients can be used to rank the rules predicting true and false verdicts $p^f(s,o), \neg p^f(s,o)$.

Explanation Context. Existing explanation approaches for KG embeddings [26, 27] mine global explanations, where the context C given as input to Alg. 1 contains a large sample of true and false statements. The latter are obtained by corrupting the true facts, so that for each fact $p(s,o)$ we also add $\{p(s',o), p(s,o')\}$ $(s \neq s', o \neq o')$. The resulting surrogate f_s approximates f's general logic when predicting p-labeled links.

A drawback of explanations based on global surrogates is that they assume that rules have always the same importance for all p-labeled predictions. Such a simplistic assumption can make explanation mining uninformative, if for example, the black box has a fine-grained behavior, i.e., it implements different logics for different regions of the KG. On those grounds, we propose to mine explanations within a *local* scope obtained by calling Alg. 1 on different sub-contexts $C' \subseteq C$ with triples that are close to each other in the latent space. These sub-contexts are obtained by applying clustering on $\boldsymbol{s} \oplus \boldsymbol{o}$, i.e., on the

latent representation of pairs s, o for true facts $p(s, o) \in C$[2]. We can also define *per-instance* contexts around a target fact $A = p(s, o)$ by calling Alg. 1 on a sub-context $C' = \{A' = p(s', o) : A' \in C\} \cup \{A' = p(s, o') : A' \in C\} \cup \{A\}$, that is, on statements that share at least one argument with A.

4 Evaluation

To answer our research questions, we study the impact of bounded atoms (**RQ1**) and locality (**RQ2**) on the fidelity of rule explanations for embedding-based link predictors through a quantitative and an anecdotal evaluation.

4.1 Experimental Setup

Datasets and Link Predictors. We resort to the benchmark datasets `fb15k-237`, `wn18rr`, and `yago3-10`, on which we trained the bilinear methods ComplEx [31] and HolE [21], and the translational approach TransE [4]. We used the implementations and data offered by the Torch-KGE library [5].

Rule Mining. We mine Horn rules with AMIE [15], a state-of-the-art rule miner for large KGs. By default, AMIE mines closed Horn rules[3] of up to 3 unbounded atoms, but it can be instructed to allow bounded atoms. AMIE does not support explicit counter-examples to estimate the precision of rules, as required by Alg. 1, hence we extended the system to support explicit false facts in the precision computation. These counter-examples were generated through a variant of Bernoulli sampling that accounts for predicate domains [33]. We use all rules making at least 2 correct predictions with a precision of at least 10% to learn the surrogate model (see Section 3).

Explanations. We compute rule-based explanations for the studied link predictors using the test instances of the experimental datasets to construct contexts C of different scopes, i.e., global, local, and per-instance as explained in Sec. 3. For each call to Alg. 1, we split C into train and test sets C_{train} and C_{test} (30%), so that we learn the explanations on C_{train} and evaluate them on C_{test}. Link predictors are mainly used for two tasks: fact classification (true vs. false) and subject/object prediction for queries $p(?, o)$ and $p(s, ?)$ where potential candidates are ranked by their score. We quantify the fidelity of our surrogate models (their ability to approximate the link predictors) for these two tasks via standard metrics, namely the ROC-AUC score and the mean reciprocal rank (MRR). The threshold θ to binarize f's scores (line 6 in Alg. 1) is chosen via logistic regression.

4.2 Results

Quantitative Evaluation. Tables 1 and 2 report the average ROC-AUC and MRR for the different explanation setups, namely unbounded vs. bounded rules

[2] $\oplus$ denotes concatenation; sub-contexts are corrupted to obtain counter-examples.
[3] These are safe rules where each variable occurs in at least 2 atoms

	ROC-AUC						S-MRR						O-MRR					
	Unbounded			Bounded			Unbounded			Bounded			Unbounded			Bounded		
	B	L	PI	G	L	PI	**B**	L	PI	G	L	PI	**B**	L	PI	G	L	PI
complex	0.71	0.68	0.64	0.93	0.93	**0.95**	0.13	0.16	0.19	0.31	0.35	**0.44**	0.97	**1.00**	0.97	0.97	0.98	0.93
transe	0.72	0.70	0.64	**0.95**	0.92	**0.95**	0.12	0.20	0.19	0.22	**0.47**	0.45	0.97	**0.99**	0.98	0.98	0.93	0.91
hole	0.66	0.63	0.60	0.98	**0.99**	**0.99**	0.08	0.16	0.22	0.27	0.36	**0.50**	0.98	0.98	0.97	0.98	**1.00**	0.97

Table 1: Fidelity on `fb15k-237`. Best performances are in bold; best locality results are underlined. The baseline **B** are global explanations with unbounded atoms. G, PI, and L stand for global, per-instance, and local explanations.

	ROC-AUC			S-MRR			O-MRR		
	G	L	PI	G	L	PI	G	L	PI
complex	0.55	0.64	**0.68**	**0.93**	0.60	0.38	0.92	0.93	**1.00**
transe	0.51	0.55	**0.69**	0.71	**0.87**	0.32	0.93	0.92	**1.00**
hole	–	0.65	**0.73**	–	**0.42**	0.38	–	**1.00**	**1.00**

(a) `wn18RR`

	ROC-AUC			S-MRR			O-MRR		
	G	L	PI	G	L	PI	G	L	PI
complex	0.55	**0.75**	0.00	**0.94**	0.39	0.17	0.87	1.00	1.00
transe	**0.66**	0.63	0.63	**0.73**	0.50	0.50	0.94	0.97	**1.00**
hole	0.71	**0.81**	0.65	**0.90**	0.38	0.26	0.93	**1.00**	0.99

(b) `yago3-10`

Table 2: Fidelity of rule-based explanations with bounded atoms.

learned on global (G), local (L), and per-instance (PI) scopes. The scores are computed by averaging the fidelity obtained for each call to Alg. 1 weighted by the size of the corresponding test set, i.e., $|C_{test}|$. We disaggregate the MRR into S-MRR and O-MRR – when the task is to predict the subject or object given the other two components.

Our baseline setting (denoted by **B**) are global unbounded rules as mined by existing approaches [7,26,27]. We highlight that we could not mine explanations with such a setting for `wn18RR` and `yago3-10` on any of the studied link predictors – not even for local or per-instance scopes. This happens because unbounded rules can only be extracted when the training KG contains very prevalent and general regularities in the interactions between the predicates. The datasets `wn18RR` and `yago3-10`, however, have much fewer predicates than `fb15k-237`: 11 and 37 for the former versus 237 for the latter. Bounded atoms also increase the coverage explanation for `fb15k-237`. While the baseline provides explanations for 18 different predicates for ComplEx on `fb15k-237`, allowing bounded rules increases the coverage to 58 predicates (HolE and TransE exhibit comparable increases). Moreover the results in Table 1 suggest that bounded atoms in rules generally increase fidelity.

It is important to remark that allowing constants in the rule atoms comes at the expense of many more, potentially noisy, rules. On `fb15k-237` with global scopes, for example, the number of unique rules mined from TransE increases from 1k to 193k. That said, only 134k of those rules get non-zero coefficients during the attribution phase – implemented via logistic regression.

We also observe that rule-based surrogates tend to be better at mimicking the link predictors for object prediction. This is explained by the nature of KG predicates, which are usually defined in a subject-oriented manner, e.g., *nationality*(*J. Biden*, *USA*) and not *hasCitizen*(*USA*, *J. Biden*). This makes subject prediction generally harder to mimic, because, e.g., it is easier to predict the nationalities of J. Biden than to predict all USA citizens. Besides, this phenomenon is corroborated by the actual performances of the link predictors. For instance, ComplEx exhibits an average S-MRR of 0.29 on `wn18RR`, whereas the average O-MRR reaches 0.46. That said, Table 2a suggests that S-MRR fidelity can still be high even in the presence of subject-oriented predicates.

When we look at the effects of locality on fidelity, we notice mixed effects. On `fb15k-237`, locality hurts ROC-AUC performance for unbounded rules and brings moderate performance gains for the MRR. The situation is different for bounded rules, for which locality boosts fidelity in most cases. These results suggest that locality and bounded rules are complementary. A similar behavior can also be observed for coverage. For example, local scopes combined with bounded rules allow mining 507k unique rules (with non-zero attribution) for 130 different predicates for ComplEx on `fb15k-237` vs. 112k rules/58 predicates and 1505 rules/62 predicates when only one of the features is enabled (the baseline mines 730 predicates covering 18 predicates). For per-instance scopes we can compute rule explanations for up to 2782 individual facts (out of 20k) covering 83 predicates (HolE on `fb15k-237`)

`fb15k-237`
(1) $place\_of\_birth(x, Chicago) \Rightarrow nationality(x, USA)$ **[TransE]**
(2) $has\_lived\_in(x, Brooklyn) \Rightarrow nationality(x, USA)$ **[ComplEx]**
(3) $profession(x, Author) \Rightarrow gender(x, M)$ **[ComplEx]**
(4) $impersonates(z, x) \wedge gender(z, y) \Rightarrow gender(x, y)$ **[ComplEx, HolE]**
(5) $country(z, y) \wedge birth\_place(x, z) \Rightarrow nationality(x, y)$
(6) $company(z, x) \wedge athlete{:}sport(z, y) \Rightarrow sport(x, y)^{\circ}$ **[HolE]**
(7) $fwc{:}club(z, x) \wedge sport(z, y) \Rightarrow sport(x, y)^{\circ}$ **[HolE]**
`yago3-10`
(8) $affiliation(x, Umeå\ IK) \Rightarrow gender(x, F)\dagger$ **[TransE, ComplEx]**
(9) $wonPrize(x, O.\ Orange\text{-}Nassau) \Rightarrow wonPrize(x, D.S.\ Medal)$ **[ComplEx]**
`wn18rr`
(10) $meronym\_mb(Insecta, x) \Rightarrow hypernym(x, Animal)\dagger$ **[ComplEx, HolE]**

Table 3: Some rule explanations. $^{\circ}$, † denote local and per-instance explanations.

Anecdotal Evaluation. Table 3 shows a few examples of rule-based explanations for our experimental link predictors. These correspond to some of the

best ranked rules according to the coefficients of the surrogate classifiers. The rules illustrate regularities preserved by the link predictors, since the body of the rules defines conditions satisfied by the facts of the KG, in contrast to the head that matches statements predicted by our black boxes (see Alg 1). Rules with bounded atoms offer legible insights about the information that the link predictors may be capturing to make predictions.

A key observation is that the different link predictors do not seem to rely on the same information – as suggested by rules (1) and (2) for ComplEx and TransE on `fb15k-237`. This is supported by the fact that among the 47 predicates for which ComplEx finds global explanations with bounded atoms, only 3 have common rules with TransE. We bring our attention to rule (3), which suggests that embeddings do reproduce the biases in the source data[4]. Recall that `fb15k-237` was mainly extracted from Wikipedia, known to have gender biases [32]. Those biases are easier to spot with rules with bounded atoms, which are a complement to more general explanations such as (4) and (5). We also highlight that local contexts can illustrate the semantics captured by the embeddings. This is exemplified by rules (6) and (7) that were learned on the same predicate but on two fact clusters. As we can see, our mining routine learned semantically equivalent rules, defined on different thematic domains, namely *athletes* and *fwc*; the latter refers to the 2010 FIFA World Cup.

5 Conclusion

We have studied the effects of specific rules with bounded atoms and local scopes on the quality of explanations for embedding-based link predictors on knowledge graphs. Our results suggest a rather positive impact on the explanation fidelity and the coverage of the explanations. Moreover, specific rules and local scopes exhibit a symbiotic relationship.

Even though rule-based explanations reflect regularities preserved by black-box link predictors, they do not shed light on causality. In this line of though, we envision to compute causal explanations that help us understand the role of the different entities, predicates, and latent components of KG embeddings in the resulting predictions. We have also planned to elaborate more on the relationship between link prediction performance and explanation fidelity, in particular at the level of the individual predicates. The source code and experimental data of our work is available at `https://gitlab.inria.fr/glatour/geebis`.

Acknowledgment. This research was supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No. 952215.

References

1. UniKER: A Unified Framework for Combining Embedding and Horn Rules for Knowledge Graph Inference. In *ICML Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020.

[4] Rule (8), on the other hand, refers to a women's football team.

2. Naser Ahmadi, Viet-Phi Huynh, Vamsi Meduri, Stefano Ortona, and Paolo Papotti. Mining Expressive Rules in Knowledge Graphs. *Journal of Data and Information Quality*, 1(1), 2019.
3. Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. Realistic Re-Evaluation of Knowledge Graph Completion Methods: An Experimental Study. In *ACM SIGMOD Conference*, 2020.
4. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*. 2013.
5. Armand Boschin. TorchKGE: Knowledge Graph Embedding in Python and PyTorch. In *International Workshop on Knowledge Graphs*, 2020.
6. Armand Boschin, Nitisha Jain, Gurami Keretchashvili, and Fabian Suchanek. Combining Embeddings and Rules for Fact Prediction. In *Int. Research School in AI in Bergen*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.
7. Ivan Sanchez Carmona and Sebastian Riedel. Extracting Interpretable Models from Matrix Factorization Models. In *International Conference on Cognitive Computation*, 2015.
8. Yang Chen, Daisy Zhe Wang, and Sean Goldberg. ScaLeKB: Scalable Learning and Inference over Large Knowledge Bases. *VLDB Journal*, 25(6), 2016.
9. Mohamed H Gad-Elrab, Daria Stepanova, Trung-Kien Tran, Heike Adel, and Gerhard Weikum. ExCut: Explainable Embedding-Based Clustering over Knowledge Graphs. In *International Semantic Web Conference*, 2020.
10. Shu Guo, Lin Li, Zhen Hui, Lingshuai Meng, Bingnan Ma, Wei Liu, Lihong Wang, Haibin Zhai, and Hong Zhang. Knowledge Graph Embedding Preserving Soft Logical Regularity. In *International Conference on Knowledge Management*, 2020.
11. Zhongni Hou, Xiaolong Jin, Zixuan Li, and Long Bai. Rule-Aware Reinforcement Learning for Knowledge Graph Reasoning. In *ACL/IJCNLP (Findings)*, 2021.
12. Zhongni Hou, Xiaolong Jin, Zixuan Li, and Long Bai. Rule-Aware Reinforcement Learning for Knowledge Graph Reasoning. In *ACL/IJCNLP (Findings)*, 2021.
13. Nitisha Jain, Trung-Kien Tran, Mohamed H. Gad-Elrab, and Daria Stepanova. Improving Knowledge Graph Embeddings with Ontological Reasoning. In *International Semantic Web Conference*, 2021.
14. Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 2022.
15. Jonathan Lajus, Luis Galárraga, and Fabian Suchanek. Fast and Exact Rule Mining with AMIE 3. In *Extended Semantic Web Conference*, 2020.
16. Ni Lao, Tom Mitchell, and William W. Cohen. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Conference on Empirical Methods in Natural Language Processing*, 2011.
17. Christian Meilicke, Patrick Betz, and Heiner Stuckenschmidt. Why a Naive Way to Combine Symbolic and Latent Knowledge Base Completion Works Surprisingly Well. In *3rd Conference on Automated Knowledge Base Construction*, 2021.
18. Christian Meilicke, Manuel Fink, Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, and Heiner Stuckenschmidt. Fine-Grained Evaluation of Rule and Embedding-Based Systems for Knowledge Graph Completion. In *International Semantic Web Conference*, pages 3–20, 2018.
19. Changping Meng, Reynold Cheng, Silviu Maniu, Pierre Senellart, and Wangda Zhang. Discovering Meta-Paths in Large Heterogeneous Information Networks. In *The Web Conference*, 2015.

20. Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.
21. Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic Embeddings of Knowledge Graphs. In *AAAI Conference on Artificial Intelligence*, 2016.
22. Maximilian Nickel and Volker Tresp. Tensor Factorization for Multi-relational Learning. In *Machine Learning and Knowledge Discovery in Databases*, 2013.
23. Georgina Peake and Jun Wang. Explanation Mining: Post Hoc Interpretability of Latent Factor Models for Recommendation Systems. In *ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, 2018.
24. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the Predictions of any Classifier. In *ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, 2016.
25. Tim Rocktäschel and Sebastian Riedel. End-to-end Differentiable Proving. In *Conference on Neural Information Processing Systems*, 2017.
26. Andrey Ruschel, Arthur Colombini Gusmão, Gustavo Padilha Polleti, and Fábio Gagliardi Cozman. Explaining Completions Produced by Embeddings of Knowledge Graphs. In *European Conf. on Symbolic and Quantitative Approaches with Uncertainty*, 2019.
27. Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. Towards Extracting Faithful and Descriptive Representations of Latent Variable Models. In *AAAI Spring Symposium on Knowledge Representation and Reasoning*, 2015.
28. Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-End Structure-Aware Convolutional Networks for Knowledge Base Completion. In *AAAI Conference on Artificial Intelligence*, 2019.
29. Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with Neural Tensor Networks for Knowledge Base Completion. In *Conference on Neural Information Processing Systems*, 2013.
30. Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*, 2019.
31. Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex Embeddings for Simple Link Prediction. In *International Conference on Machine Learning*, 2016.
32. Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the Glass Ceiling: Gender Asymmetries in Wikipedia. *EPJ Data Science*, 5:1–24, 2016.
33. Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge Graph Embedding by Translating on Hyperplanes. *AAAI Conference on Artificial Intelligence*, 28(1), 2014.
34. Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *International Conference on Learning Representations*, 2015.
35. Wen Zhang, Bibek Paudel, Liang Wang, Jiaoyan Chen, Hai Zhu, Wei Zhang, Abraham Bernstein, and Huajun Chen. Iteratively Learning Embeddings and Rules for Knowledge Graph Reasoning. In *The Web Conference*, 2019.
36. Yongqi Zhang, Quanming Yao, and Lei Chen. Efficient, Simple and Automated Negative Sampling for Knowledge Graph Embedding, 2020.