Knowledge Graph Embedding with Logical Consistency

Jianfeng Du[1], Kunxun Qi[1], and Yuming Shen[1]

Guangdong University of Foreign Studies, Guangzhou 510006, China
jfdu@gdufs.edu.cn

Abstract. Existing methods for knowledge graph embedding do not ensure the high-rank triples predicted by themselves to be as consistent as possible with the logical background which is made up of a knowledge graph and a logical theory. Users must take great effort to filter consistent triples before adding new triples to the knowledge graph. To alleviate users' burden, we propose an approach to enhancing existing embedding-based methods to encode logical consistency into the learnt distributed representation for the knowledge graph, enforcing high-rank new triples as consistent as possible. To evaluate this approach, four knowledge graphs with logical theories are constructed from the four great classical masterpieces of Chinese literature. Experimental results on these datasets show that our approach is able to guarantee high-rank triples as consistent as possible while preserving a comparable performance as baseline methods in link prediction and triple classification.

1 Introduction

Knowledge graph has become popular in knowledge representation nowadays. A knowledge graph is a directed graph with vertices labeled by entities and edges labeled by relations. It is represented as a set of triples of the form $\langle h, r, t\rangle$, where h is the *head entity* (simply *head*), r the *relation* and t the *tail entity* (simply *tail*). Although the current knowledge graphs have many triples, they are far from completeness. Traditional logic-based methods for knowledge graph completion become intractable as knowledge graphs evolve into a large scale.

To handle large knowledge graphs, embedding-based methods have emerged and become prevalent due to their high efficiency and scalability. These methods introduce the distributed representation for entities and relations and encode them into a continuous vector space, then the truth degrees of any new triples can be estimated by simple numerical calculation. In the scenario of knowledge graph completion, all new triples are ranked in the descending order of truth degrees computed by an embedding-based method, therein high-rank triples having passed the verification of domain experts are added to the knowledge graph.

Logical consistency is a prerequisite for new triples to be usable in knowledge graph completion. A new triple cannot be added to a knowledge graph if it is inconsistent with the logical background which consists of the knowledge graph and a logical theory formalizing the domain knowledge. The logical theory can often be expressed by both *datalog rules* and *constraints*, where datalog rules are used to infer implicit triples and constraints are used to prevent conflicting triples from being present. For example, suppose $\langle \mathsf{John}, \mathsf{father}, \mathsf{Tom} \rangle$ (meaning *John is the farther of Tom*) is an existing triple, $\langle x, \mathsf{father}, y \rangle \rightarrow \langle x, \mathsf{type}, \mathsf{Man} \rangle$ (meaning *x is a man if x is the father of someone y*) and $\langle x, \mathsf{mother}, y \rangle \rightarrow \langle x, \mathsf{type}, \mathsf{Woman} \rangle$ (meaning *x is a woman if x is the mother of someone y*) are two datalog rules, and $\langle x, \mathsf{type}, \mathsf{Man} \rangle \wedge \langle x, \mathsf{type}, \mathsf{Woman} \rangle \rightarrow \bot$ (meaning *anyone x cannot be both a man and a woman*) is a constraint, respectively, in the logical background. Then the new triple $\langle \mathsf{John}, \mathsf{mother}, \mathsf{Tom} \rangle$ (meaning *John is the mother of Tom*) is inconsistent with the logical background and should not be added to the knowledge graph that underlies the logical background.

Although logical consistency is important, existing embedding-based methods for knowledge graph completion hardly address logical consistency in learning the distributed representation. Some methods such as the translational distance methods [3, 2, 24, 9, 15, 11, 14, 12, 13, 26] and the semantic matching methods [10, 20, 1, 27, 21, 17, 16] completely ignore the logical theory. Others [19, 23, 25, 4, 7, 5, 8] consider datalog rules but ignore constraints in the logical theory. None of them makes high-rank new triples as consistent as possible with the logical background. Users must take great effort to throw away inconsistent triples before adding new triples to the knowledge graph. Although automatic tools for checking logical consistency can be applied to filter high-rank triples, this post-processing is still time consuming. It is wiser to encode logical consistency into the distributed representation to make high-rank triples as consistent as possible.

In this paper, we propose an approach to enhancing existing embedding-based methods to encode logical consistency into the learnt distributed representation. The key idea is to adapt knowledge graph embedding into an optimization problem that minimizes the sum of the global margin-based loss function in translational distance methods and a loss function on a portion of negative triples that are inconsistent with the logical background, where inconsistent triples are computed from the logical background and selected by adaptive relation-specific thresholds. We do not use the complete set of inconsistent triples in the optimal objective function for consideration of efficiency and scalability.

Since there is no benchmark knowledge graph coming with a logical theory including constraints, we construct new knowledge graphs and the corresponding logical theories from a domain that we are familiar with, which is about character relationships in the four great classical masterpieces of Chinese literature, to evaluate our approach. Experimental results demonstrate that our approach is able to guarantee the high-rank new triples as consistent as possible while it also preserves a comparable performance as baseline methods in the two traditional prediction tasks namely link prediction and triple classification.

The main contributions of this work are two-fold:

(1) We propose knowledge graphs and logical theories from the four great classical masterpieces of Chinese literature. These datasets give complex theories on the domain of character relationships and are suitable for evaluating hybrid approaches that combine distributed representation learning with logic inference.

(2) We propose an effective and efficient approach to enhancing embedding-based methods to guarantee high-rank new triples as consistent as possible with the logical background while keeping comparable scores on traditional metrics.

2 Preliminaries

2.1 Knowledge Graph Embedding

Knowledge graph embedding (KGE) encodes a knowledge graph into a continuous vector space to support a variety of prediction tasks. There are roughly two main categories for KGE methods, namely the translational distance methods and the semantic matching methods.

A translational distance method measures the truth degree of a triple by the distance between the head and the tail, usually after a translation carried out by the relation. TransE [2] is a pioneer translational distance method which defines translation directly on entity vectors. Most subsequent methods define translation on projection of entity vectors. To name a few, TransH [24] defines projection on relation-specific hyperplanes, whereas TransR [15] defines projection by relation-specific matrices. All these methods uniformly learn vectors by minimizing a global margin-based loss function via a loss function for triples.

The loss function for a triple $\langle h, r, t\rangle$, written $\mathsf{loss}_r(h, t)$, can be defined as $||f_r(l_h) + l_r - f_r(l_t)||_{L_1}$ or $||f_r(l_h) + l_r - f_r(l_t)||_{L_2}$, where l_x denotes $(x_1, \ldots, x_n)$ which is the vector representation for x, $||l_x||_{L_1}$ is the L1-norm of l_x defined as $\sum_{i=1}^{n} |x_i|$, $||l_x||_{L_2}$ is the L2-norm of l_x defined as $\sqrt{\sum_{i=1}^{n} x_i^2}$, and $f_r(\cdot)$ is a relation-specific projection function which maps an entity vector into another one. For example, $f_r(l_e)$ is defined as l_e in TransE, whereas it is defined as $l_e - w_r^{\mathrm{T}} l_e w_r$ for w_r an r-specific normal vector in TransH.

The global margin-based loss function to be minimized is defined over the set $\mathcal{G}$ of training triples (namely positive triples) and a set $\overline{\mathcal{G}}$ of negative triples which is disjoint with $\mathcal{G}$ and constructed from $\mathcal{G}$ by randomly corrupting triples in either heads or tails. By introducing the margin γ as a hyper-parameter, the global margin-based loss function to be minimized is given by

$$\sum_{\langle h,r,t\rangle \in \mathcal{G}} \sum_{\langle h',r,t'\rangle \in \overline{\mathcal{G}}} \max(0, \gamma + \mathsf{loss}_r(h, t) - \mathsf{loss}_r(h', t')). \tag{1}$$

A semantic matching method measures truth degrees of triples by matching the latent semantics for entities and relations in their vector space. It often employs a neural network to learn the distributed representation. Due to the diversity of network structures, semantic matching methods often have not a uniform optimal objective function. Nevertheless, a state-of-the-art method Analogy [16] has been proved to be generalized from several semantic matching methods including DistMult [27], HolE [17] and ComplEx [21]. Simply speaking, Analogy minimizes the energy function $\mathbb{E}_{h,r,t,y \in \mathcal{D}} - \log_2 \sigma(y \cdot l_h^{\mathrm{T}} B_r l_t)$ on the data distribution $\mathcal{D}$ constructed from the training triples $\langle h, r, t\rangle$, where $y = +1$ for training triples (namely positive triples) and $y = -1$ for negative triples, l_h and

l_t are respectively the vector representations for h and t, B_r is an $m \times m$ almost-diagonal matrix with $n < m$ real scalars on the diagonal, and $\sigma(\cdot)$ denotes the sigmoid function. DistMult (resp. ComplEx) can be treated as a special case of Analogy such that $n = m$ (resp. $n = 0$). Moreover, Analogy can be reduced to HolE by setting B_r as a certain circulant matrix without real scalars.

High-rank triples that have the top-k highest truth degrees (i.e. lowest loss values or energies) are often used as candidates in knowledge graph completion.

2.2 Logical Consistency

First-order logic is a traditional and popular approach to knowledge representation. A logical theory expressed by first-order logic is a set of rules R of the form $\forall \boldsymbol{x}\ (\phi(\boldsymbol{x}) \rightarrow \exists \boldsymbol{y}\ \varphi(\boldsymbol{x}, \boldsymbol{y}))$, where $\phi(\boldsymbol{x})$ is a conjunction of atoms on the universally quantified variables $\boldsymbol{x}$, and $\varphi(\boldsymbol{x}, \boldsymbol{y})$ is a disjunction of atoms on both universally quantified variables $\boldsymbol{x}$ and existentially quantified variables $\boldsymbol{y}$. The part of R at the left (resp. right) of $\rightarrow$ is called the *body* (resp. *head*) of R. By $\text{body}(R)$ (resp. $\text{head}(R)$) we denote the set of atoms in the body (resp. head) of R. If the head of R has no atoms, R is also called a *constraint* and the empty head is written as $\perp$. If the head of R has a single atom without existentially quantified variables, R is also called a *datalog* rule. We omit $\forall \boldsymbol{x}$ and simply write constraints as $\phi(\boldsymbol{x}) \rightarrow \perp$ and datalog rules as $\phi(\boldsymbol{x}) \rightarrow \varphi(\boldsymbol{x})$.

Existing KGE methods that take logical theories into account [19, 23, 25, 4, 7, 5, 8] focus on datalog rules but ignore constraints. In this work we study KGE under a *Horn theory* which is a logical theory consisting of both datalog rules and constraints. A *model* of a Horn theory can be defined as a set S of *ground atoms* (which can be treated as a set of triples) such that (1) $\text{body}(R)\,\theta \in S$ implies $\text{head}(R)\,\theta \in S$ for any datalog rule R and any ground substitution θ for $\text{var}(R)$ in the theory, and (2) $\text{body}(R)\,\theta \notin S$ for any constraint R and any ground substitution θ for $\text{var}(R)$ in the theory, where $\text{var}(R)$ denotes the set of variables in R. A constraint R is said to be *entailed* by a Horn theory if every model of the theory is also a model of $\{R\}$. A *logical background* is made up of a knowledge graph (namely a set of triples) and a Horn theory. A model of a logical background $\mathcal{B}$ that is made up of a knowledge graph $\mathcal{G}$ and a Horn theory $\mathcal{T}$ is a model of $\mathcal{T}$ that is also a superset of $\mathcal{G}$. A logical background *is* said to be *consistent* if it has a model. A triple is said to be *logically consistent* (simply *consistent*) with a consistent logical background $\mathcal{B}$ if adding the triple to $\mathcal{B}$ keeps $\mathcal{B}$ consistent. A triple is said to be *entailed* by a consistent logical background $\mathcal{B}$ if the triple is in the unique least model of $\mathcal{B}$.

3 Knowledge Graph Embedding under a Horn Theory

Given a logical background with a Horn theory, new triples that can be added to the knowledge graph underlying the background should be consistent with the logical background. We focus on the task of finding consistent high-rank new triples. It is unwise to filter consistent triples after applying an existing

KGE method to compute the truth degrees of new triples, since it is rather time-consuming — it needs to perform consistency checking for every new triple separately, where consistency checking in Horn theories is polynomial-time complete. It is wiser to encode logical consistency into the distributed representation to guarantee consistency for new triples as possible, so that the costly consistency checking need not to be performed after new triples are computed.

We propose an approach to enhancing translational distance KGE methods under a logical background. It is also applied to any semantic matching KGE method after the energy function is treated as a loss function and the optimal objective function is rewritten to a global margin-based loss function. The enhancements are two-fold. On one hand, to take datalog rules into account, we treat the set of entailed triples (which is actually the least model of the logical background) as the set of positive triples. On the other hand, to take constraints into account, we add to the optimal objective function a loss function on certain negative triples that are inconsistent with the logical background, where inconsistent triples are computed from the logical background and selected by adaptive relation-specific thresholds. While the former enhancement has been proposed in existing work [7, 5], the latter enhancement is novel. These two enhancements can be formalized to the following optimal objective function:

$$\sum_{\langle h,r,t\rangle\in\mathcal{G}^+}\sum_{\langle h',r,t'\rangle\in\overline{\mathcal{G}^+}}\max(0,\gamma+\mathsf{loss}_r(h,t)-\mathsf{loss}_r(h',t')) + \sum_{\langle h,r,t\rangle\in\mathcal{G}^-}\max(0,\sigma_r-\mathsf{loss}_r(h,t)), \tag{2}$$

where $\mathcal{G}^+$ is the unique least model of the logical background, $\overline{\mathcal{G}^+}$ is constructed from $\mathcal{G}^+$ by randomly corrupting triples on either heads or tails, $\mathcal{G}^-$ is the set of negative triples inconsistent with the logical background, and σ_r is the r-specific threshold that classifies positive or negative triples on the relation r; i.e., $\langle h,r,t\rangle$ is a positive triple if $\mathsf{loss}_r(h,t)<\sigma_r$, or a negative triple otherwise.

More details are provided as follows. Let $\mathcal{B}$ denote logical background with a knowledge graph $\mathcal{G}$ and a Horn theory $\mathcal{T}$, where the set of datalog rules in $\mathcal{T}$ is $\mathcal{T}_1$ and the set of constraints in $\mathcal{T}$ is $\mathcal{T}_2$. Following [5] $\mathcal{G}^+$ is computed as the least fixpoint of $\mathcal{G}^{(t)}$, where $\mathcal{G}^{(t)}=\mathcal{G}^{(t-1)}\cup\bigcup\{\text{head}(R)\,\theta\mid R\in\mathcal{T}_1,\text{body}(R)\,\theta\in\mathcal{G}^{(t-1)}\}$ for $t>0$, and $\mathcal{G}^{(0)}=\mathcal{G}$. Following [24] $\overline{\mathcal{G}^+}$ is constructed from $\mathcal{G}^+$ by the *bern* strategy e.g. by randomly corrupting triples on either heads or tails with different probabilities proportional to the relation-specific frequency of heads or tails. Let R_{-1} denote the set of constraints modified from the constraint R by deleting one body atom. For efficiency, $\mathcal{G}^-$ can be approximated as $\bigcup_{R\in\mathcal{T}_2}\bigcup_{R'\in R_{-1}}\{\langle h,r,t\rangle\mid \text{body}(R')\,\theta\in\mathcal{G}^+,\text{body}(R)\,\theta=\{\langle h,r,t\rangle\}\cup\text{body}(R')\,\theta\}$. Such a $\mathcal{G}^-$ is a subset of the complete set of inconsistent triples, since for every triple $\langle h,r,t\rangle\in\mathcal{G}^-$, adding $\langle h,r,t\rangle$ to $\mathcal{G}^+$ will instantiate the body of some constraint in $\mathcal{T}_2$ to a subset of $\mathcal{G}^+$, which means that it will render $\mathcal{B}$ inconsistent and thus $\langle h,r,t\rangle$ is an inconsistent triple. We can make $\mathcal{G}^-$ closer to the complete set of inconsistent triples by adding to $\mathcal{T}_2$ more constraints that are entailed by $\mathcal{T}$. Finally, with the learnt vectors for entities and relations up to now, following [20] σ_r is set as the value that maximizes $\Sigma_{\langle h,r,t\rangle\in\mathcal{G}^+} I(\mathsf{loss}_r(h,t) < \sigma_r) + \Sigma_{\langle h',r,t'\rangle\in\overline{\mathcal{G}^+}} I(\mathsf{loss}_r(h',t') \geq \sigma_r)$, namely the number of correctly classified triples on r, where $I(C)$ is the indicator function such that $I(C) = 1$ if C is true or $I(C) = 0$ otherwise.

Following existing translational distance methods, the above optimal objective function is minimized by stochastic gradient descent. To guarantee efficiency and convergence, the set of relation-specific thresholds $\{\sigma_r\}$ is updated for every k training rounds, where k can be set from tens to hundreds. For the first k rounds, the loss values of inconsistent triples are ignored; i.e., the optimal objective function is reduced to the standard global margin-based loss function. Afterwards, the set of relation-specific thresholds is computed for the first time. In other iterations, since there have been relation-specific thresholds, the optimal objective function is recovered as given in Formula (2).

4 Experimental Evaluation

4.1 Implementation

We enhanced two translational distance methods (TransE [2] and TransH [24]) and three semantic matching methods (DistMult [27], ComplEx [21] and Analogy [16]) by our proposed approach. We chose these methods because they are the most efficient methods in their respective categories. To uniformly use a global margin-based loss function as the optimal objective function for semantic matching methods, the loss function for triples $\langle h,r,t\rangle$ in Analogy is defined as $\mathsf{loss}_r(h,t) = -l_h^{\mathrm{T}} B_r l_t$. This loss function is restricted by setting n real scalars (resp. 0 real scalar) on the diagonal of B_r in DistMult (resp. ComplEx), where n is the dimension of entity vectors. By DistMult$^{\mathsf{tr}}$, ComplEx$^{\mathsf{tr}}$ and Analogy$^{\mathsf{tr}}$ we denote the variants of DistMult, ComplEx and Analogy that use Formula (1) as the optimal objective function, respectively. Moreover, by X-lc we denote the method enhanced from the baseline method X by our proposed approach.

We implemented all the above methods with multi-threads in Java, using stochastic gradient descent with fixed mini-batch size 1, and evaluated them in the RapidMiner platform[1] to ensure all methods to be compared in the same environment. For the baseline methods we replaced the input training set $\mathcal{G}$ in Formula (1) with the triple set $\mathcal{G}^+$ entailed by the union of the logical thoery and the training set. This implementation of baseline methods has been shown to be effective in improving the predictive performance [7, 5]. For the implementation of Analogy$^{\mathsf{tr}}$ and Analogy$^{\mathsf{tr}}$-lc, we fixed the number of real scalars on the diagonal of B_r as $\frac{n}{2}$ for n the dimension of entity vectors.

To tune the hyper-parameters in our implemented methods, we focus on three prediction tasks, namely link prediction, triple classification and consistency checking for high-rank triples, where the last task is studied for the first time.

Link prediction, originated from [2], is a benchmark task in the field of KGE which aims to predict the missing head or the missing tail in a triple. We used

[1] `https://www.rapidminer.com/`

the conventional metrics Mean Reciprocal Rank (MRR) and Hits@k (where $k = 1, 10$) to evaluate the performance in the test set. Since a corrupted triple in the evaluation is actually correct if it is entailed by the logical background, we computed the ranks of triples without counting any entailed triples.

Triple classification, originated from [20], is another benchmark task aiming to judge whether a triple is correct or not. We used the conventional metric Classification Accuracy (CA) to evaluate the performance in the test set. This metric is computed by first determining the relation-specific thresholds on loss values of positive triples in the training set and then calculating the ratio of correctly classified positive triples and negative triples in the test set, where one negative triple is generated from every positive triple.

Consistency checking for high-rank triples aims to judge whether a high-rank triple is consistent with the logical background or not. We introduced the metrics Prec@k (where $k = 10, 100, 1000$) to evaluate the performance. Prec@k is defined as the proportion of triples that are consistent with the logical background among the top-k *new* triples that have the smallest loss values, where the set of new triples is defined as the set of triples which are constructed from existing entities and relations but are not entailed by the logical background.

In our experiments, we fixed the number of training rounds to 2000 in all the above methods and the interval for updating the relation-specific thresholds to 100 rounds in all enhanced methods. Moreover, we selected the learning rate among $\{0.001, 0.005, 0.01, 0.05, 0.1\}$, the margin among $\{0.1, 0.15, 0.2, 0.5, 1, 2\}$, the dimension of entity vectors among $\{20, 50, 100, 200\}$, and the dissimilarity measure as either L1-norm or L2-norm. Since our goal is to learn an embedding model that works well for all the aforementioned tasks, we determined the optimal hyper-parameters that achieve the highest geometric mean of three metrics Hits@1, CA and Prec@10 by five-fold cross-validation on the training set.

4.2 Dataset Construction

Existing benchmark datasets for KGE do not have corresponding logical theories with constraints. It is hard to add adequate constraints to these datasets since they belong to general domains using our unfamiliar relations. Thus we constructed new knowledge graphs and the corresponding logical theories from a domain that we are familiar with. The domain is about character relationships in the four great classical masterpieces of Chinese literature, namely Dream of the Red Chamber (DRC), Journey to the West (JW), Outlaws of the Marsh (OM), and Romance of the Three Kingdoms (RTK). We manually built the logical theories for these four masterpieces separately in Protege[2], a well-known ontology editor. These theories model character relationships which are originally expressed in OWL 2 [6], a W3C recommended language for modeling ontologies, and then translated to Horn theories by standard syntactic transformation. We collected triples on character relationships from e-books, yielding four knowledge graphs each of which corresponds to one of these masterpieces.

[2] `https://protege.stanford.edu/`

Table 1. Knowledge graphs and their corresponding logical theories

KG	#rel	#ent	#all	#train	#ent-train	#test	#datalog	#const	#ent-const
DRC	48	392	382	344	5,377	323	110	107	737
JW	31	104	109	99	1,136	154	94	30	154
OM	49	156	201	181	1,975	339	111	47	282
RTK	51	123	153	138	3,123	342	142	63	389

Note: #rel/#ent/#all are respectively the number of relations/entities/triples in the knowledge graph, #train/#test are respectively the cardinality of the training set/test set, #ent-train is the number of triples entailed by the union of the logical theory and the training set, #datalog/#const/#ent-const are respectively the number of datalog rules/constraints/entailed constraints in the logical theory.

We split a knowledge graph into a training set and a seed set by 9:1. Initially the training set is the complete knowledge graph and the seed set is empty. Then the seed set is enlarged by randomly picking out triples from the training set one by one, until the number of triples in the seed set reaches one-tenth of the cardinality of the knowledge graph, where every triple picked out is composed by entities in the remainder of the training set but is not entailed by the union of the logical theory and the remainder of the training set. The test set is defined as the set of triples entailed by the union of the logical theory and the seed set.

Recall that we approximated the set $\mathcal{G}^-$ of negative triples in Formula (2). When more entailed constraints are used, the approximation will be closer to the exact result. To find more entailed constraints, we treated constraints in the logical theory as conjunctive queries and applied the REQUIEM tool[3] [18] to compute all constraints entailed by the $\mathcal{ELHIO}$ fragment of the logical theory, where $\mathcal{ELHIO}$ is a description logic underlying OWL 2 [6].

Table 1 reports the statistics about our constructed datasets. These datasets and our implemented systems are available at OpenKG.CN[4].

4.3 Results on Link Prediction and Triple Classification

Table 2 reports the evaluation results on link prediction and triple classification for all baseline methods X and their enhanced methods X-lc, under their corresponding optimal hyper-parameters determined by five-fold cross-validation on the training set. It can be seen that, for all the datasets the metric score of an enhanced method can be higher or lower than the metric score of the corresponding baseline method, while the difference between them is rather small.

Table 3 further summarizes the comparison results between baseline and enhanced methods for all datasets. It can be seen that the mean metric score of the baseline methods is slightly higher than that of their enhanced methods, but the difference is trivial as the p-value in t-test for accepting the null hypothesis

[3] `http://www.cs.ox.ac.uk/projects/requiem/`
[4] `http://www.openkg.cn/tool/kge-wlc`

Table 2. Evaluation results on link prediction and triple classification

KG	DRC				JW			
Metric	MRR	Hits@1	Hits@10	CA	MRR	Hits@1	Hits@10	CA
TransE	0.453	0.432	0.483	0.698	0.811	0.786	0.838	**0.873**
TransE-lc	0.442	0.429	0.457	0.700	0.758	0.714	0.831	0.860
TransH	0.453	0.430	0.488	0.697	0.815	0.792	0.844	0.870
TransH-lc	0.440	0.430	0.454	**0.704**	0.812	0.773	**0.899**	0.864
DistMult^{tr}	0.452	0.432	**0.489**	0.676	0.817	0.795	0.844	0.870
DistMult^{tr}-lc	**0.461**	**0.441**	0.486	0.683	0.800	0.763	0.838	0.821
ComplEx^{tr}	0.453	**0.441**	0.474	0.693	**0.821**	**0.805**	0.851	0.864
ComplEx^{tr}-lc	0.432	0.415	0.450	0.684	0.812	0.799	0.831	0.870
Analogy^{tr}	0.441	0.427	0.457	0.697	0.790	0.773	0.812	0.864
Analogy^{tr}-lc	0.430	0.416	0.440	**0.704**	0.814	0.799	0.831	0.860
KG	OM				RTK			
Metric	MRR	Hits@1	Hits@10	CA	MRR	Hits@1	Hits@10	CA
TransE	0.842	0.811	**0.912**	0.844	0.869	0.849	**0.905**	**0.930**
TransE-lc	0.845	0.814	**0.912**	0.844	0.868	0.848	0.901	0.925
TransH	0.848	0.813	0.909	0.845	0.863	0.842	0.893	0.920
TransH-lc	0.847	0.813	0.910	0.892	0.863	0.842	0.898	0.923
DistMult^{tr}	**0.857**	**0.835**	0.886	0.886	0.855	0.830	0.893	0.918
DistMult^{tr}-lc	0.828	0.788	0.883	**0.907**	0.858	0.842	0.895	0.909
ComplEx^{tr}	0.827	0.799	0.882	0.897	0.860	0.845	0.886	0.923
ComplEx^{tr}-lc	0.837	0.811	0.878	0.875	**0.874**	**0.860**	0.890	0.928
Analogy^{tr}	0.820	0.802	0.845	0.894	0.855	0.841	0.870	0.918
Analogy^{tr}-lc	0.815	0.798	0.850	0.883	0.860	0.845	0.890	0.917

Table 3. The mean±std metric score and the p-value in t-test for all datasets

Metric	MRR	Hits@1	Hits@10	CA
Baseline methods	0.740±0.173	0.719±0.171	0.773±0.177	0.839±0.090
Enhanced methods	0.735±0.176	0.712±0.172	0.771±0.188	0.838±0.089
p-value	0.92484	0.89759	0.97512	0.97042

that the based methods have the same mean metric score as their enhanced methods have is close to 1. This implies that the enhanced methods achieve comparable performance as baseline methods in traditional prediction tasks.

4.4 Results on Consistency Checking for High-rank Triples

Table 4 reports the evaluation results on consistency checking for high-rank triples for all baseline methods X and their enhanced methods X-lc, under their corresponding optimal hyper-parameters. It can be seen that, for all the datasets the metric score of an enhanced method is very high (always ≥ 0.9) and is consistently not lower than the metric score of the corresponding baseline method.

Table 4. Evaluation results on consistency checking for high-rank triples

KG	DRC			JW			OM			RTK		
Prec@k	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
TransE	0.4	0.48	0.438	0.9	0.90	0.491	**1.0**	0.95	0.842	**1.0**	0.99	0.981
TransE-lc	**1.0**	**1.00**	**1.000**	**1.0**	0.96	0.927	**1.0**	**1.00**	0.984	**1.0**	**1.00**	0.985
TransH	0.5	0.47	0.434	0.8	0.88	0.586	**1.0**	0.98	0.969	**1.0**	0.98	0.981
TransH-lc	**1.0**	**1.00**	0.968	**1.0**	**0.99**	0.913	**1.0**	**1.00**	0.979	**1.0**	0.99	0.985
DistMult^{tr}	0.5	0.37	0.670	0.1	0.80	0.953	0.7	0.92	0.979	0.6	0.60	0.549
DistMult^{tr}-lc	**1.0**	**1.00**	**1.000**	0.9	0.91	0.941	**1.0**	**1.00**	**1.000**	**1.0**	**1.00**	**1.000**
ComplEx^{tr}	0.7	0.65	0.735	0.6	0.65	0.898	0.8	0.83	0.972	0.9	0.72	0.645
ComplEx^{tr}-lc	**1.0**	**1.00**	0.999	0.9	0.90	**0.966**	**1.0**	0.97	0.982	**1.0**	**1.00**	0.949
Analogy^{tr}	0.8	0.77	0.802	0.5	0.82	0.947	0.9	0.92	0.981	0.8	0.64	0.620
Analogy^{tr}-lc	**1.0**	**1.00**	**1.000**	**1.0**	0.95	0.963	**1.0**	**1.00**	**1.000**	**1.0**	**1.00**	**1.000**

Table 5. The mean±std metric score and the p-value in t-test for all datasets

Metric	Prec@10	Prec@100	Prec@1000
Baseline methods	0.725±0.240	0.766±0.187	0.774±0.202
Enhanced methods	0.990±0.031	0.984±0.031	0.977±0.026
p-value	0.00002	0.00001	0.00007

Table 5 further summarizes the comparison results between baseline and enhanced methods for all datasets. It can be seen that the mean metric score of the baseline methods is clearly lower than that of their enhanced methods, while this difference is significant since the p-value in t-test for accepting the null hypothesis that the baseline methods have the same mean metric score as their enhanced methods have is nearly 0.

5 Related Work

As mentioned in Subsection 2.1, translational distance methods and semantic matching methods constitute the two main categories for KGE methods. The well-known translational distance methods include Structured Embedding (SE) [3], TransE [2], TransH [24], TransR [15], TransD [11], PTransE [14], TranSparse [12], TransA [13], KG2E [9], TransG [26], etc. Among the above methods, except for KG2E and TransG that use distribution functions to represent entities and relations, others represent entities and relations by vectors and employ Formula (1) as the optimal objective function. On the other hand, the well-known semantic matching methods include Latent Factor Model (LFM) [10], Single Layer Model (SLM) [20], Neural Tensor Network (NTN) [20], Semantic Matching Energy (SME) [1], DistMult [27], ComplEx [21], HolE [17], Analogy [16], etc. These methods exploit neural networks with various structures to learn a distributed representation. Among these methods Analogy can be treated as a generalized model for DistMult, ComplEx and HolE [16]. We refer the interested readers to [22] for a thorough review of existing KGE methods.

The most related work to us is about combining KGE with logic inference. Wang et al. [23] and Wei et al. [25] tried to combine embedding models with datalog rules for KGE. But in their work, rules are modeled separately from embedding models and would not help to learn a more predictive distributed representation. Rocktäschel et al. [19], Guo et al. [7] and Du et al. [5] proposed several joint learning paradigms that encode the inference of implicit triples into KGE, where the inference is guided by datalog rules. To avoid the costly propositionalization of datalog rules, Demeester et al. [4] further proposed a method for injecting lifted datalog rules to KGE. The above join models make a one-time injection of hard rules, taking them as either additional training instances [19, 7, 5] or regularization terms [4]. To make the best of background information, Guo et al. [8] proposed a paradigm for interactively injecting soft rules to KGE, which learns simultaneously from labeled triples, unlabeled triples and soft rules with various confidence levels in an iterative manner. However, the above work considers only datalog rules that are used to infer implicit triples, but ignores constraints that prevent new triples from being present.

6 Conclusions and Future Work

To ensure the high-rank triples predicted by an embedding-based method to be as consistent as possible with the logical background, which consists of a knowledge graph and a Horn theory, we have proposed an approach to enhancing the method to encode logical consistency into the learnt distributed representation. To evaluate this approach, we also constructed four knowledge graphs and their corresponding Horn theories from the four great classical masterpieces of Chinese literature. Experimental results on these datasets demonstrated the efficacy of our proposed approach. For future work, we plan to study improved methods to achieve higher metric scores in link prediction and triple classification, as well as to consider more expressive logical theories beyond Horn theories to guarantee logical consistency in knowledge graph embedding.

Acknowledgements. This work was partly supported by National Natural Science Foundation of China (61375056), Science and Technology Program of Guangzhou (201804010496), and Scientific Research Innovation Team in Department of Education of Guangdong Province (2017KCXTD013).

References

1. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: A semantic matching energy function for learning with multi-relational data - Application to word-sense disambiguation. Machine Learning **94**(2), 233–259 (2014)
2. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS. pp. 2787–2795 (2013)
3. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. In: AAAI (2011)
4. Demeester, T., Rocktäschel, T., Riedel, S.: Lifted rule injection for relation embeddings. In: EMNLP. pp. 1389–1399 (2016)

5. Du, J., Qi, K., Wan, H., Peng, B., Lu, S., Shen, Y.: Enhancing knowledge graph embedding from a logical perspective. In: JIST. pp. 232–247 (2017)
6. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P.F., Sattler, U.: OWL 2: The next step for OWL. Journal of Web Semantics **6**(4), 309–322 (2008)
7. Guo, S., Wang, Q., Wang, L., Wang, B., Guo, L.: Jointly embedding knowledge graphs and logical rules. In: EMNLP. pp. 192–202 (2016)
8. Guo, S., Wang, Q., Wang, L., Wang, B., Guo, L.: Knowledge graph embedding with iterative guidance from soft rules. In: AAAI. pp. 4816–4823 (2018)
9. He, S., Liu, K., Ji, G., Zhao, J.: Learning to represent knowledge graphs with gaussian embedding. In: CIKM. pp. 623–632 (2015)
10. Jenatton, R., Roux, N.L., Bordes, A., Obozinski, G.: A latent factor model for highly multi-relational data. In: NIPS. pp. 3176–3184 (2012)
11. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: ACL. pp. 687–696 (2015)
12. Ji, G., Liu, K., He, S., Zhao, J.: Knowledge graph completion with adaptive sparse transfer matrix. In: AAAI. pp. 985–991 (2016)
13. Jia, Y., Wang, Y., Lin, H., Jin, X., Cheng, X.: Locally adaptive translation for knowledge graph embedding. In: AAAI. pp. 992–998 (2016)
14. Lin, Y., Liu, Z., Luan, H., Sun, M., Rao, S., Liu, S.: Modeling relation paths for representation learning of knowledge bases. In: EMNLP. pp. 705–714 (2015)
15. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: AAAI. pp. 2181–2187 (2015)
16. Liu, H., Wu, Y., Yang, Y.: Analogical inference for multi-relational embeddings. In: ICML. pp. 2168–2178 (2017)
17. Nickel, M., Rosasco, L., Poggio, T.A.: Holographic embeddings of knowledge graphs. In: AAAI. pp. 1955–1961 (2016)
18. Pérez-Urbina, H., Motik, B., Horrocks, I.: Tractable query answering and rewriting under description logic constraints. Journal of Applied Logic **8**(2), 186–209 (2010)
19. Rocktäschel, T., Singh, S., Riedel, S.: Injecting logical background knowledge into embeddings for relation extraction. In: NAACL. pp. 1119–1129 (2015)
20. Socher, R., Chen, D., Manning, C.D., Ng, A.Y.: Reasoning with neural tensor networks for knowledge base completion. In: NIPS. pp. 926–934 (2013)
21. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: ICML. pp. 2071–2080 (2016)
22. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017)
23. Wang, Q., Wang, B., Guo, L.: Knowledge base completion using embeddings and rules. In: IJCAI. pp. 1859–1866 (2015)
24. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI. pp. 1112–1119 (2014)
25. Wei, Z., Zhao, J., Liu, K., Qi, Z., Sun, Z., Tian, G.: Large-scale knowledge base completion: Inferring via grounding network sampling over selected instances. In: CIKM. pp. 1331–1340 (2015)
26. Xiao, H., Huang, M., Zhu, X.: TransG: A generative model for knowledge graph embedding. In: ACL. pp. 992–998 (2016)
27. Yang, B., Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. CoRR **abs/1412.6575** (2014), `http://arxiv.org/abs/1412.6575`