Loss-aware Pattern Inference: A Correction on the Wrongly Claimed Limitations of Embedding Models

Mojtaba Nayyeri[1,2], Chengjin Xu[1], Yadollah Yaghoobzadeh[3], Sahar Vahdati[2], Mirza Mohtashim Alam[1,2], Hamed Shariat Yazdi[1], and Jens Lehmann[1,4]

[1] University of Bonn, Bonn, Germany
nayyeri,shariat@cs.uni-bonn.de
[2] InfAI Lab, Dresden, Germany
vahdati,mohtasim@infai.org
[3] Microsoft
Yaghoobzadeh@m
[4] Fraunhofer IAIS, Dresden, Germany
jens.lehmann@iais.fraunhofer.de

Abstract. Knowledge graph embedding models (KGEs) are actively utilized in many of the AI-based tasks, especially link prediction. Despite achieving high performances, one of the crucial aspects of KGEs is their capability of inferring relational patterns, such as symmetry, antisymmetry, inversion, and composition. Among the many reasons, the inference capability of embedding models is highly affected by the used loss function. However, most of the existing models failed to consider this aspect in their inference capabilities. In this paper, we show that disregarding loss functions results in inaccurate or even wrong interpretation from the capability of the models. We provide deep theoretical investigations of the already exiting KGE models on the example of the TransE model. To the best of our knowledge, so far, this has not been comprehensively investigated. We show that by a proper selection of the loss function for training a KGE e.g., TransE, the main inference limitations are mitigated. The provided theories together with the experimental results confirm the importance of loss functions for training KGE models and improving their performance.

1 Introduction

Recent years witnessed a great attention on the topic of knowledge graph embedding (KGE) models such that they rapidly became one of the state-of-the-art methods for Link Prediction. One of the primary KGE models is TransE which gained a lot of attention due to its simplicity and high performance. Some follow up models tried to improve TransE in terms of encoding relation types such as 1-many, or certain relation patterns such as reflexive, and symmetric [**?**,**?**,**?**]. While the community got into a paradigm of proposing new embedding models by (only) focusing on the score function and competing on decimal improvements of the results, the actual cause that was rooted in the loss function, remained overlooked. Although, in a separate track, several loss functions have been proposed [**?**], its role in studying the capability of KGE models in presence of relational patterns have been majorly ignored. This neglected fact resulted in inaccurate or even wrong interpretations of the model capability.

Our investigations showed that, this problem originates in the initial assumption used for model capability proofs. To formally show this, let (h, r, t) be a positive triple in a KG where h, t are the entities and r is the relation between them which are to be embedded in $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ (vector representation). This fact started from the TransE model, and continued by TransH, TransR, SimplE [**?**,**?**,**?**], where the assumption of $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ was used in evaluating the limitations of the models in encoding of patterns. This has caused a boom of new KGE models addressing the claimed limitations by proposing only new score functions. On the example of symmetric relations (e.g. brotherOf), this means that by enforcing a relation r to be symmetric ($\mathbf{t} + \mathbf{r} \approx \mathbf{h}$), the relation embedding is then enforced to be $\mathbf{r} = \mathbf{0}$. In this way, all of the corresponding vectors for entities that are related to each other via r relation will be equal. This had the risk of being interpreted as "model disability" in encoding symmetric patterns. While this problem was interpreted as model limitation, it was caused by incompatible equality assumption in the loss functions. We consider this as off-track argument followed by many of the KGE proposals overlooking the root cause. Here, we cover some of these works. In [**?**], additional limitations of TransE, FTransE [**?**], STransE [**?**], TransH and TransR are addressed which are listed here: (i) if the models encode a reflexive relation r, they automatically encode symmetric; (ii) if the models encode a reflexive relation r, they automatically encode transitive and; (iii) if entity h_1 has relation r with every entity in $\Delta \in \mathcal{E}$ and entity h_2 has relation r with one of entities in Δ, then h_2 must have the relation r with every entity in Δ. All of these limitations are justified by the initial assumption which was never fulfilled as they were incompatible with the utilized loss functions. Thus, these proposed approaches are only shedding the light on the underlying score functions.

Even a recent KGE model namely RotatE followed the same problem, which is a highly valuable work, however, it also claims that TransE is not capable of encoding symmetric patterns (Table 2 in [**?**]), considering the same assumption (i.e. $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$). The equality assumption is satisfied when the loss function enforces $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \approx 0$. The claimed disability has been argued to be caused by the score function of TransE. However, none of the existing loss functions (i.e. Margin Ranking Loss [**?**] and Adversarial Los [**?**]) hold this assumption during the optimization process, rather such losses take $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| \leq \gamma_1$ where γ_1 is upper-bound of positive scores. Therefore, most of the identified limitations of the existing KGEs and the proposed solutions for them have been based on an assumption that was not fulfilled by any of the existing loss functions. We re-studied the reported limitations of the TransE model employing the "appropriate" assumption compatible with the used loss function, as we shall see in the body of this paper, TransE is capable of encoding symmetric patterns. Although we highlight that ignoring loss functions caused inaccurate results on studying the limitations of TransE, it is generalziable for other existing KGE models as well as for the models yet to come (if this paradigm continuous). The impact of work is in blocking such a continuous misinterpretation of the inference capability of KGE models influenced by ignored loss functions with a long standing inappropriate assumption. Moreover, our theoretical finding is consistent with the recent experimental studies [**?**] which highlight that old models perform as well as the recent state-of-the-art models if they are trained with the same setting (using the same boosting techniques).

In summary, our main contributions are the following:

- We show that the different loss functions enforce different upper-bounds and lower-bounds for the scores of positive and negative samples respectively.
- We illustrate that the existing theories corresponding to the limitations of translation-based models are inaccurate since they only consider the score functions. We prove theoretically and later experimentally that the selection of loss functions is critical and can mitigate the main limitations.
- Using symmetric relation patterns, we obtain a proper upper-bound of positive triples score to enable encoding of symmetric patterns.
- We prove that applying translation in the complex space gives a more powerful model while efficiency in memory and time is preserved.

2 Related Works

Most of the previous work, majorly investigate the capability of the translation-based embedding models solely considering the formulation of the score functions. Accordingly, in this section, we review the score functions of TransE and its variants.

The score of **TransE** [**?**] is initially defined as $f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$. In order to overcome the problems of TransE in encoding of relational patterns, **TransH** [**?**] was proposed where the score function was modified as $f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|$. In this way, each entity (**e**) is projected to a relation space ($\mathbf{e}_\perp = \mathbf{e} - \mathbf{w}_r\mathbf{e}\mathbf{w}_r^T$). Using this score function, the TransH model reported itself to be capable of encoding reflexive, one-to-many, many-to-one and many-to-many relations. This effort was done while the identified problem of TransE being incapable of encoding relational patterns was not valid. However, other works [**?**,**?**] started to build up on top of TransH addressing its problems. For example, encoding reflexive pattern leads to undesired encoding of both symmetric and transitive relations [**?**].

TransR [**?**] was then proposed with a new score function that projects each entity (**e**) to the relation space by using a matrix provided for each relation ($\mathbf{e}_\perp = \mathbf{e}\mathbf{M}_r$, $\mathbf{M}_r \in R^{d_e \times d_r}$). The chain of new score functions has continued with **TransD** [**?**] which provides two vectors for each individual entities and relations ($\mathbf{h}, \mathbf{h}_p, \mathbf{r}, \mathbf{r}_p, \mathbf{t}, \mathbf{t}_p$). Head and tail entities are projected using the following matrices: $\mathbf{M}_{rh} = \mathbf{r}_p^T\mathbf{h}_p + \mathbf{I}^{m\times n}, \mathbf{M}_{rt} = \mathbf{r}_p^T\mathbf{t}_p + \mathbf{I}^{m\times n}$. The score function of TransD is similar to the score of TransH.

Recently, the **RotatE** [**?**] model has been proposed to address encoding of relational patterns with a new score function. It rotates the head to the tail entity using relation in the Complex space. Using constraints on the norm of entity vectors, the model is reformed to TransE. The scoring function of RotatE is $f_r(h,t) = \|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|$, where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in C^d$, and $\circ$ is element-wise product. RotatE obtains the state-of-the-art results using very big embedding dimension (1000) and a lot of negative samples (1000). **TorusE** [**?**] fixes the problem of regularization in TransE by applying translation on a compact Lie group. The model has several variants including mapping from Torus to Complex space. In this case, the model is regarded as a special case of RotatE [**?**] applying rotation instead of translation in the Complex space. According to [**?**], TorusE is not defined on the entire Complex space. Therefore, it has less representation capacity. TorusE needs a very big embedding dimension (10000 as reported in [**?**]) which is

a limitation. All of these state-of-the-art models only focus on proposing a new score function based on an assumption that was not valid for the used losses.

3 Loss-aware Pattern Inference

In this section, we first introduce the wrongly interpreted limitations of the embedding models especially TransE and its follow ups (Section 3.1). Then, the limitations are re-investigated in the light of *score* and *loss* functions where we show that the corresponding theoretical proofs are inaccurate because the effect of loss function is ignored (Section 3.2). So, we propose new theories and prove that each of the limitations of TransE is resolvable by revising either the *scoring* function which the community continued with high effort or re-investigating the *loss* with regard to the limitations in the base assumption (our work).

3.1 Wrong Interpretations of KGE Models Presented as Limitations

Here we discuss the wrong interpretations of the KGE models which was reported in several literature that ended up to be presented as their limitations [**?**,**?**,**?**,**?**]. We focus on the reported limitations (L) of translation-based embedding models and their encoding capabilities for relation patterns (e.g. reflexive, symmetric) as following:

L1 TransE cannot encode reflexive relations when the relation vector is non-zero [**?**].
L2 TransE cannot encode a relation r which is neither reflexive nor irreflexive. To see that, if TransE encodes the relation r, we have $\mathbf{h}_1 + \mathbf{r} = \mathbf{h}_1$ and $\mathbf{h}_2 + \mathbf{r} \neq \mathbf{h}_2$, resulting $\mathbf{r} = \mathbf{0}$, $\mathbf{r} \neq \mathbf{0}$ which is a contradiction [**?**].
L3 TransE cannot encode symmetric relation when $\mathbf{r} \neq \mathbf{0}$. If r is symmetric, then: $\mathbf{h} + \mathbf{r} = \mathbf{t}$ and $\mathbf{t} + \mathbf{r} = \mathbf{h}$. Thus, $\mathbf{r} = \mathbf{0}$ and all entities appeared in head or tail of triples will have the same vector [**?**].
L4 If r is reflexive on $\Delta \in \mathcal{E}$, where $\mathcal{E}$ is the set of all entities in the KG, then r must also be symmetric [**?**].
L5 If r is reflexive on $\Delta \in \mathcal{E}$, r must also be transitive [**?**].
L6 If entity h_1 has relation r with every entity in Δ and entity h_2 has relation r with one of entities in Δ, then h_2 must have relation r with every entity in Δ [**?**].

Limitations 3 to 5 have been reported for TransE, however they are genelaizable for all the follow up models. Limitations 4 to 6 have been reported for TransE, FTransE, STransE, TransH and TransR.

3.2 Re-investigation of the Reported Limitations

Here, we aim at analyzing the limitations of TransE (in real and complex spaces) by considering the effect of both *score* and *loss* functions. A loss function determines the score boundary within which a triple is positive or negative. A KGE model considers a triple (h, r, t) to be positive if its score is in a region of truth and a triple (h', r, t') to be negative if its score is in a region of falsity.

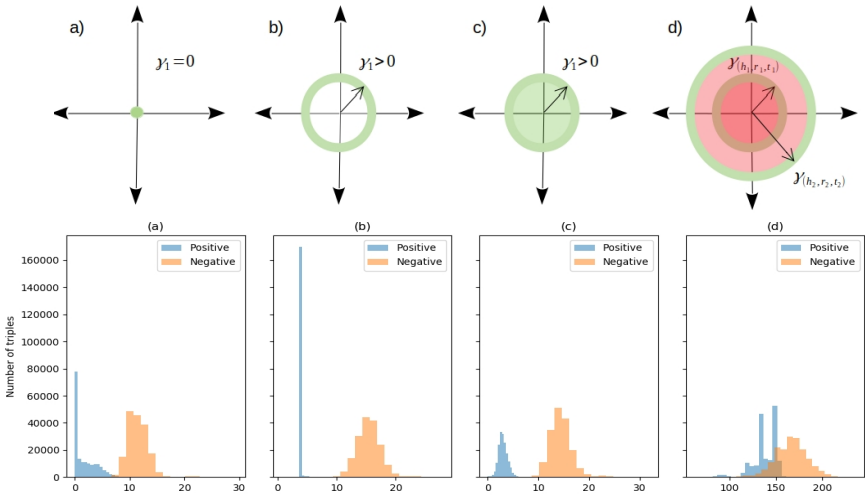

Fig. 1: **Top:** Visualization of truth region (positive) according to Table 1. The residual vector ϵ, (a) becomes $\mathbf{0}$, (b) lies on the border of a sphere with radius γ_1, (c) lies inside of a sphere with radius γ_1, and (d) $\epsilon_{(h_1,r_1,t_1)}$ lies inside of a sphere with radius $\gamma_{(h_1,r_1,t_1)}$. **Bottom:** The histogram of scores when TransE is trained on WordNet using the losses of Equation 2 ($\gamma_1 = 0$), 2 ($\gamma_1 = 4$), 4 ($\gamma_1 = 4$) and 6 ($\gamma = 6$) respectively. Each histogram is the approximation of the corresponding conditions (a)-(d).

Such a boundary enforces an assumption through which the capability of embedding models in encoding relation pattern is investigated. For instance, in TransE with score function of $f_r(h,t) = \|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$, the already used assumption for boundary of positive and negative samples are $f_r(h,t) = 0$ and $f_r(h',t') > 0$, respectively. However, this can not be fulfilled (or even approximated) by the considered loss functions of the state-of-the-art models (e.g. margin ranking loss [**?**] and RotatE loss [**?**]).

To re-investigate and address the reported limitations, we propose four new conditions (Table-1) where a triple can be considered positive or negative by the score function. This is done by considering the defined thresholds for upper- and lower-bounds (decision boundary) in the scores of both positive and negative triples. We show that these conditions can be approximated by designing appropriate loss functions. In this regards, we adapt four loss functions based on [**?**] and propose compatible conditions for each of them (see Table 1). To better comprehend this, we illustrated the conditions in Figure 1. The condition (a) indicates that a triple is positive if $\mathbf{h} + \mathbf{r} = \mathbf{t}$ holds. It means that the length of *residual vector* i.e. $\epsilon = \mathbf{h} + \mathbf{r} - \mathbf{t}$, is zero. It is the most strict condition that expresses the extent to which a triple is positive. Authors in [**?**,**?**] consider this condition to prove their theories as well as the limitation of TransE in the encoding of symmetric relations. However, the employed loss function fails to approximate (a), rather it fulfills condition (c) which results a void limitation in that setting. The condition (b) considers a triple to be positive if its residual vector lies on a hypersphere with radius γ_1. It is less restrictive than (a) which only considers a point in the vector space

Table 1: Region of truth and falsity.

Condition	Positive	Negative	$\gamma_1, \gamma_2 \in R$
(a)	$f_r(h,t) = \gamma_1,$	$f_r(h',t') \geq \gamma_2$	$\gamma_1 = 0,\ \gamma_2 > 0$
(b)	$f_r(h,t) = \gamma_1$	$f_r(h',t') \geq \gamma_2$	$\gamma_2 > \gamma_1 > 0$
(c)	$f_r(h,t) \leq \gamma_1$	$f_r(h',t') \geq \gamma_2$	$\gamma_2 > \gamma_1 > 0$
(d)	$f_r(h,t) \leq \gamma_{1(h,r,t)}$	$f_r(h',t') \geq \gamma_{2(h,r,t)}$	$\gamma_{2(h,r,t)} > \gamma_{1(h,r,t)} > 0$

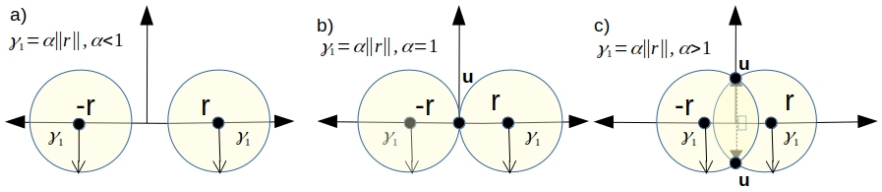

Fig. 2: Condition for encoding symmetric relation: (a) when $\alpha < 1$, the model cannot encode it. (b) when $\alpha = 1$, the intersection of two hyperspheres is a point. **u = 0** means embedding vectors of all the entities should be equal. Symmetric cannot be encoded. (c) when $\alpha > 1$, symmetric can be encoded as there are more than one point in the intersection of two hyperspheres.

to express the positiveness of a triple. The optimization problem that approximates the conditions (a) ($\gamma_1 = 0$) and (b) ($\gamma_1 > 0$) is as follows:

$$\begin{cases} \min_{\xi_{h,t}} \sum_{(h,r,t)\in S^+} {\xi_{h,t}}^2 \\ s.t.\ f_r(h,t) = \gamma_1,\ (h,r,t) \in S^+ \\ f_r(h',t') \geq \gamma_2 - \xi_{h,t},\ (h',r,t') \in S^- \\ \xi_{h,t} \geq 0 \end{cases} \tag{1}$$

where S^+, S^- are the sets of positive and negative samples. $\xi_{h,t}$ are slack variables to reduce the effect of noise in negative samples. One loss function that approximates the conditions (a) and (b) is as follows where for case (a), we set $\gamma_1 = 0$ and for case (b) we set $\gamma_1 > 0$ in the formula.

$$\mathcal{L}_{a|b} = \sum_{(h,r,t)\in S^+} \Big(\lambda_1 \|f_r(h,t) - \gamma_1\| + \sum_{(h',r,t')\in S^-_{(h,r,t)}} \lambda_2 \max(\gamma_2 - f_r(h',t'), 0) \Big). \tag{2}$$

The condition (c) considers a triple to be positive if its residual vector is inside a hypersphere of radius γ_1. The optimization problem that approximates the condition (c) is:

$$\begin{cases} \min_{\xi_{h,t}} \sum_{(h,r,t)\in S^+} {\xi_{h,t}}^2 \\ f_r(h,t) \leq \gamma_1,\ (h,r,t) \in S^+ \\ f_r(h',t') \geq \gamma_2 - \xi_{h,t},\ (h',r,t') \in S^- \\ \xi_{h,t} \geq 0 \end{cases} \tag{3}$$

The loss function that approximates the condition (c) is:

$$\mathcal{L}_c = \sum_{(h,r,t)\in S^+} \Big(\lambda_1 \max(f_r(h,t) - \gamma_1, 0) \; + \sum_{(h',r,t')\in S^-_{(h,r,t)}} \lambda_2 \; \max(\gamma_2 - f_r(h',t'), 0) \Big). \tag{4}$$

Remark: The loss function which is defined in [**?**] is slightly different from the loss in Equation4. The former loss slides the margin while the latter fixes the margin by inclusion of a lower-bound for the score of negative triples. Both losses put an upper-bound for scores of positive triples. Apart from the loss 4, the RotatE loss [**?**] also approximates the condition (c). The formulation of the RotatE loss is as follows:

$$\mathcal{L}_c^{RotatE} = -\sum_{(h,r,t)\in S^+} \Big(\log \sigma(\gamma - f_r(h,t)) + \sum_{(h',r,t')\in S^-_{(h,r,t)}} \log \sigma(f_r(h',t') - \gamma) \Big). \tag{5}$$

The condition (d) is similar to (c), but it provides different γ_1, γ_2 for each triple. Using (d), there is a triple-specific region of truth for each positive triple (h,r,t) and its corresponding negative triple (h',r,t'). Margin ranking loss [**?**] approximates (d). Defining $[x]_+ = \max(0,x)$, the loss is:

$$\mathcal{L}_d = \sum \sum [f_r(h,t) + \gamma - f_r(h',t')]_+ \, . \tag{6}$$

To re-investigate the limitations, we must assume that the relation vectors do not get zero values otherwise we will have the same embedding for head and tail which is undesirable. Here, the conditions (a) to (d) are presented by the following theorems.

Theorem T1. *(Addressing L1)*: TransE (real and complex) cannot infer a reflexive relation pattern with a non-zero relation vector under (a). However, under (b-d), TransE (real and complex) can infer reflexive pattern.

Theorem T2. *(Addressing L2)*: (i) TransE (complex) can infer a relation which is neither reflexive nor irreflexive under (b-d). (ii) TransE (real) cannot infer a relation which is neither reflexive nor irreflexive under (a-d).

Theorem T3. *(Addressing L3)*: (i) TransE (complex) can infer symmetric relations under (a-d). (ii) TransE (real) cannot infer symmetric relations under (a) with non-zero vector for relation. (iii) TransE (real) can infer a symmetric relation under (b-d). *Proof:* Here, due to space problems we only prove (iii) as a representative for other proofs.

Under (b), for TransE we have $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| = \gamma_1$ and $\|\mathbf{t} + \mathbf{r} - \mathbf{h}\| = \gamma_1$. The necessity condition for encoding symmetric relation is $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\| = \|\mathbf{t} + \mathbf{r} - \mathbf{h}\|$. This implies $\|\mathbf{h}\| \cos(\theta_{h,r}) = \|\mathbf{t}\| \cos(\theta_{t,r})$. Let $\mathbf{h} - \mathbf{t} = \mathbf{u}$, by definition we have $\|\mathbf{u} + \mathbf{r}\| = \gamma_1, \|\mathbf{u} - \mathbf{r}\| = \gamma_1$. Now let $\gamma_1 = \alpha\|r\|$, we have:

$$\begin{cases} \|\mathbf{u}\|^2 + (1-\alpha^2)\|\mathbf{r}\|^2 = -2\langle \mathbf{u}, \mathbf{r} \rangle \\ \|\mathbf{u}\|^2 + (1-\alpha^2)\|\mathbf{r}\|^2 = 2\langle \mathbf{u}, \mathbf{r} \rangle \end{cases} \tag{7}$$

Therefore, there is: $\|\mathbf{u}\|^2 + (1-\alpha^2)\|\mathbf{r}\|^2 = -(\|\mathbf{u}\|^2 + (1-\alpha^2)\|\mathbf{r}\|^2)$, which can be written as $\|\mathbf{u}\|^2 = (\alpha^2 - 1)\|\mathbf{r}\|^2$. To avoid contradiction, we must have $\alpha > 1$. Once $\alpha > 1$, we have $\cos(\theta_{u,r}) = \pi/2$. Therefore, TransE can encode symmetric relation with condition (b), when $\gamma_1 = \alpha\|r\|$ and $\alpha > 1$. Figure 2 shows different conditions

for encoding symmetric relation. Conditions (c-d) are directly resulted from (b), as it is subsumed by (c) and (d). That completes the proof.

Theorem T4. *(Addressing L4)*: For TransE (real and complex) (i) Limitation L4 holds under (a). (ii) Limitation L4 is not valid under (b-d).

Theorem T5. *(Addressing L5)*: For TransE (real and complEx) (i) Limitation L5 holds under (a). (ii) Limitation L5 holds is not valid under (b-d).

Theorem T6. *(Addressing L6)*: For TransE (real and complex), (i) Limitation L6 holds under (a). (ii) Limitation L6 is not valid under (b-d).

4 Experiments and Evaluations

In this section, we evaluate the performance of TransE in real and complex spaces with different loss functions used for theoretical analysis of the limitations. The experiments are done for a link prediction task with the aim of completing the triple $(h, r, ?)$ or $(?, r, t)$ by predicting the missing entities for h or t. Filtered Mean Rank (MR), Mean Reciprocal Rank (MRR) and Hits@10 are the evaluation metrics [**?**,**?**]. We used two evaluation dataset namely FB15K-237 [**?**] and WN18RR [**?**].

4.1 Experimental Setup.

We implement TransE (real and complex) with the losses 2, 4 and 6 in PyTorch. Adagrad is used as an optimizer and 100 mini-batches have been generated in each iteration. The hyperparameter corresponding to the score function is embedding dimension d. We add slack variables to the losses 2 and 4 to have soft margin as in [**?**]. The loss 4 is rewritten as follows:

$$\min_{\xi^r_{h,t}} \sum_{(h,r,t)\in S^+} \big(\lambda_0 {\xi^r_{h,t}}^2 + \lambda_1 \max(f_r(h,t) - \gamma_1, 0) + \lambda_2 \sum_{(h,r,t)\in S^-_{h',r,t'}} \max(\gamma_2 - f_r(h^{'}, t^{'}) - \xi^r_{h,t}, 0)\big). \tag{8}$$

4.2 Results and Discussion

In this part, we compare TransE (real and complex) and RotatE trained by using the losses 2 (condition (a),(b)), 4 (condition (c)) and the RotatE loss (condition (c)). For FB15K-237, we set the embedding dimension to 300 and the number of negative samples to 256. For WN18RR, we set the embedding dimension and the number of negative samples to 300 and 250 respectively. We additionally use adversarial negative sampling technique from [**?**] that we have applied for all the models.

Analysis of the results: Table 2 presents a comparison of TransE (real and complex) and RotatE trained by different losses. TransE in Equation $2(\gamma_1 = 0)$ is trained by using the loss in Equation 2 when $\gamma_1 = 0$. TransE in Equation $2(\gamma_1 > 0)$ refers to the TransE model which is trained by using the loss Equation 2 when γ_1 is a non-zero positive value. The TransE model which is trained by the losses in Equation 4 and the RotatE loss (i.e., $\mathcal{L}_c^{RotatE}$) are denoted by TransE4 and TransE$\mathcal{L}_c^{RotatE}$ respectively. Similar notations are considered for TransE (complex) and RotatE when they are trained by

	FB15K-237			WN18RR		
	MR	MRR	Hits@10	MR	MRR	Hits@10
TransE2 ($\gamma_1 = 0$)	222	27.4	45.7	3014	19.3	47.4
TransE2 ($\gamma_1 > 0$)	198	31.3	50.5	3942	21.4	50.3
TransE4	181	32.3	52.1	3451	23.5	53.9
TransE$\mathcal{L}_c^{RotatE}$	179	32.5	51.9	3594	23.3	53.6
TransE (complex)2 ($\gamma_1 = 0$)	213	28.5	47.3	3014	31.2	49.5
TransE (complex)2 ($\gamma_1 > 0$)	194	31.9	50.8	3942	41.3	50.8
TransE (complex)4	177	32.8	52.1	3435	44.3	55.0
TransE (complex)$\mathcal{L}_c^{RotatE}$	176	32.7	51.9	3537	44.2	54.7
RotatE4	194	33.0	52.0	3806	47.8	56.9
RotatE$\mathcal{L}_c^{RotatE}$	196	33.0	51.8	3943	47.3	56.5

Table 2: Link prediction results. Rows 1-4: TransE trained using condition (a), (b) (with loss 2 (c)(with the loss 4) and (c)(with the RotatE loss) with no injected relation patterns. Rows 5-8 TransE (complex) trained using condition (a), (b), (c)(with the loss 4) and (c)(with the RotatE loss) with no injected relation patterns. Rows 9-10: RotatE trained using condition (c)(with the loss 4) and (c)(with the RotatE loss) with no injected relation patterns.

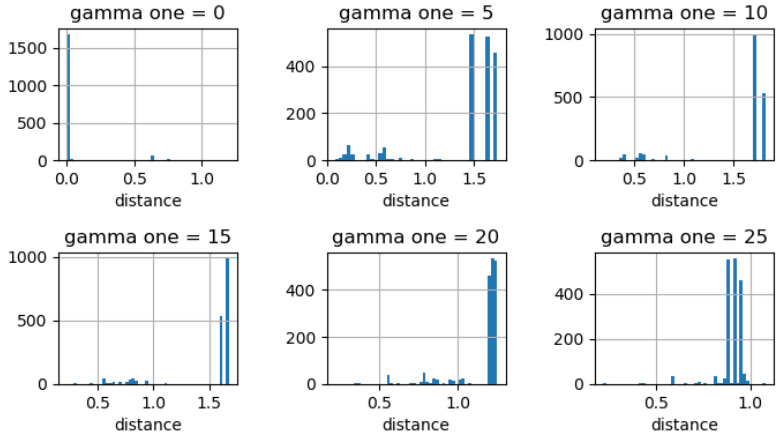

Fig. 3: Histogram of $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ for reflexive triple ($\mathbf{h} = \mathbf{t}$) per different γ_1.

using different loss functions. The loss 2 with $\gamma_1 = 0$ approximates the condition (a), and the approximation is done with condition (b) for $\gamma_1 > 0$.

The condition (c) can be approximated by using the loss in Equation 4 and the RotatE loss (i.e., $\mathcal{L}_c^{RotatE}$). However, the loss Equation 4 provides a better separation for positive and the negative samples than the RotatE loss. According to the Table 2, the loss

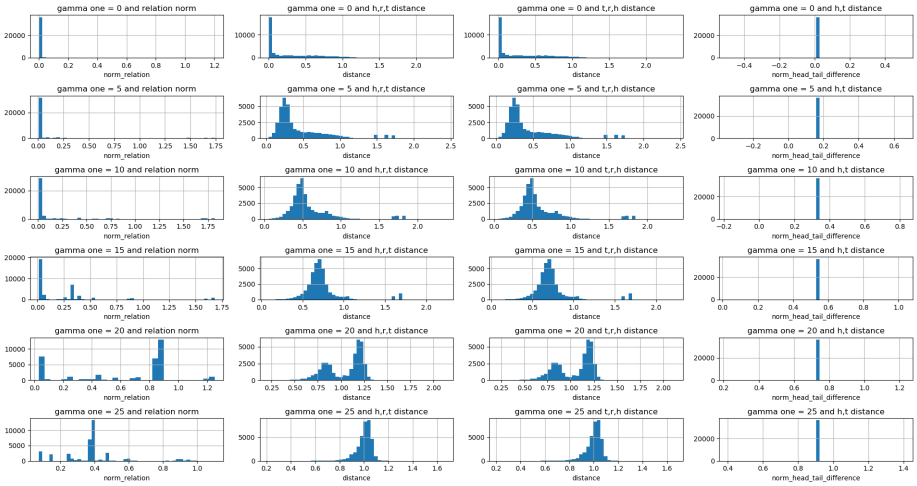

Fig. 4: Scores of symmetric triple for different gamma.

4 obtains a better performance than the other losses in each class of the studied models. It is consistent with our theories indicating that the condition (c) is less restrictive.

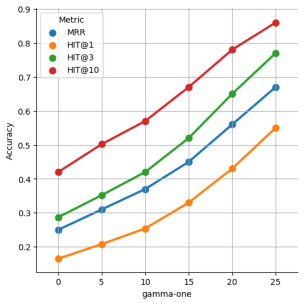

Fig. 5: Accuracy with respect to evaluation metric.

Although we only focus on translation-based KGEs, the theories can be generalized to different models including the RotatE model. We see that the loss in Equation 4 improves the performance of RotatE. Regarding the Table 2, the loss in Equation 2 ($\gamma_1 = 0$) gets the worst results. It confirms our theories that with the condition (a), most of the limitations are held. However, with the condition (c), the limitations no longer exist. Previously, there have not been any loss that approximates the condition (a). However, most of the theories presented above corresponding to the main limitations of the translation-based class of embedding models (L1-L6) have been proven using the condition (a) while the used loss didn't approximate the condition. Therefore, in all of the previous works the theories and experimental justifications have not been accurate.

4.3 Further Analysis of Theories

In Theorem Th1-Th6, we proved that most of the claimed limitations about TransE are inaccurate or even incorrect due to wrong assumptions not to be fulfilled by the used losses. More concretely, the claimed limitations were not rooted in the formulation of score function of TransE. Even worse, the claimed limitations were not also rooted in the loss function. The claimed limitations are analytically derived by using

wrong assumption which is not fulfilled by the used loss function. Here we visualize the histogram of distance function ($\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$) to encode symmetric relation using various $\gamma_1 = \{0, 5, 10, 15, 20, 25\}$. While most of proofs corresponding to the limitations of TransE have been done with condition (a) ($\gamma_1 = 0$), the used loss function (margin ranking loss does not fulfill that condition (see Figure 1)). Figure 3 visualizes the histogram of distance when the relation r is reflexive.

In the case of reflexive relation, distance $\|\mathbf{e}+\mathbf{r}-\mathbf{e}\| = \|\mathbf{r}\|$ is the norm of relation. According to this results, when $\gamma_1 = 0$, the norm of relation must be zero to consider the triple as positive. Therefore, with non-zero relation, all triples will be recognized as negative sample. In the case of $\|\mathbf{r}\| = 0$, for all other triples in the form of (h, r, t) where $h \neq t$, the embeddings of head and tail must be equal which is undesired. However, from the Figure 3 (the cases of $\gamma_1 \neq 0$, we can see that most of reflexive relations are non-zero. Moreover, we have $\|\mathbf{e}+\mathbf{r}-\mathbf{e}\| = \|\mathbf{r}\| \leq \gamma_1 \neq 0$. Therefore, the triples (e, r, e) are learned as positive by the model while the embeddings of relation is not zero. This confirms that the claim about TransE not being capable of encoding reflexive relation is no longer valid when margin ranking loss is used. Figure 4 shows that only when $\gamma_1 = 0$, the embedding of relation becomes zero, which addresses the limitations of L3. According to the last figure of the first row, when $\gamma_1 = 0$, the embeddings of head and tail becomes equal. Therefore, with non-zero relation vector, the TransE model cannot encode symmetric relation with condition (a). However, from the last row of Figure 4, with a bigger value for upper-bound of positive triples $\gamma_1 = 25$, the embedding of relation is not zero and when all triples (h, r, t) and their symmetric (t, r, h) are learned as positive (encoding symmetric by TransE) based on the second and third columns of the last row in Figure 4. Moreover, from the last sub-figure of last row, we see that embedding of head and tail are different. This shows symmetric relation is properly learned by the model. From Figure **??**, we observe that by increasing γ_1, the accuracy of the model increases when learning is done on symmetric patterns. As a conclusion, the mentioned limitations of TransE do not exist because there none of the previous loss functions fulfill the condition (a).

5 Conclusion

In this paper, we re-investigated the main limitations of Translation-based embedding models from two aspects: *score* and *loss*. We showed that different loss functions enforce different boundaries for triple scores, affecting the limitations of embedding models in encoding relation patterns such as symmetric. Therefore, the existing theories corresponding to the limitations of the KGE models are inaccurate because the effect of loss functions has been ignored. Accordingly, we presented new theories about the limitations by consideration of the effect of score and loss functions. The TransE model (in both real and Complex space) is trained by using various loss functions on standard datasets. According to the experiments, TransE in complex space with appropriate loss function significantly outperformed other existing translation-based embedding models. It got competitive performance with the other embedding models while it is more efficient in time and memory. Beside the performance-related improvements, the main impact of our work is the correction it provides between the initial assumption and the

used loss function by most of the already existing embedding models. The objective is to influence the future embedding models and shed light on the effect of loss functions.
Acknowledgements. We acknowledge the support of the EU projects TAILOR (GA 952215), Cleopatra (GA 812997), the BmBF project MLwin, the EU Horizon 2020 grant 809965, and ScaDS.AI (01/S18026A-F).