Learning Knowledge Embeddings by Combining Limit-based Scoring Loss

Xiaofei Zhou
Institute of Information Engineering, Chinese Academy of Sciences, China & University of Chinese Academy of Sciences, School of Cyber Security, China
zhouxiaofei@iie.ac.cn

Qiannan Zhu
Institute of Information Engineering, Chinese Academy of Sciences, China & University of Chinese Academy of Sciences, School of Cyber Security, China
zhuqiannan@iie.ac.cn

Ping Liu
Institute of Information Engineering, Chinese Academy of Sciences, China & University of Chinese Academy of Sciences, School of Cyber Security, China
liuping@iie.ac.cn

Li Guo*
Institute of Information Engineering, Chinese Academy of Sciences, China & University of Chinese Academy of Sciences, School of Cyber Security, China
guoli@iie.ac.cn

ABSTRACT

In knowledge graph embedding models, the margin-based ranking loss as the common loss function is usually used to encourage discrimination between golden triplets and incorrect triplets, which has proved effective in many translation-based models for knowledge graph embedding. However, we find that the loss function cannot ensure the fact that the scoring of correct triplets must be low enough to fulfill the translation. In this paper, we present a limit-based scoring loss to provide lower scoring of a golden triplet, and then to extend two basic translation models TransE and TransH, separately to TransE-RS and TransH-RS by combining limit-based scoring loss with margin-based ranking loss. Both the presented models have low complexities of parameters benefiting for application on large scale graphs. In experiments, we evaluate our models on two typical tasks including triplet classification and link prediction, and also analyze the scoring distributions of positive and negative triplets by different models. Experimental results show that the introduced limit-based scoring loss is effective to improve the capacities of knowledge graph embedding.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Knowledge representation and reasoning**; **Machine learning**; • **Information systems** → Entity relationship models;

KEYWORDS

Embedding; Knowledge graph; Representation learning; Distributed representation

1 INTRODUCTION

Knowledge graph as an effective way to represent knowledge has made a contribution to artificial intelligence and knowledge management [17, 22, 25, 32]. For example, various available large-scale knowledge graphs such as WordNet [21], Freebase [2], GeneOntology [1], NELL [8] and Yago [23] have become very important resources to support intelligence application and knowledge management [3, 9, 13, 19, 20, 26, 27, 30, 31, 36, 37].

Knowledge graphs are multi-relational directed graphs composed of entities as nodes and relations as edges, in which a triplet (head, relation, tail) denoted as (h, r, t) represents a relationship r from head h to tail t. The aim of a knowledge graph completion is to predict relations and determine specific-relation type between entities. To fulfill the aim, many of the current methods under supervision of the existing triplets to learn knowledge embeddings show strong feasibility and robustness [4–7, 11, 12, 15, 16, 18, 24, 31, 33–35], such as Structured Embedding [4], Semantic Matching Energy Model [5, 6], Neural Tensor Network Model [28], TransE [7] and TransH [33] etc.

Among these methods, translation based models are promising to encode entities as low dimensional embeddings and relationships between entities as translation vectors. Usually a relation-dependent translation scoring function, such as $f_r(h,t) = \|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$, is defined to measure the correctness of a triplet (h, r, t) in the embedding space. $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$ when (h, r, t) holds, while $\mathbf{h'}+\mathbf{r}$ should be far away from $\mathbf{t'}$ for a corrupted triplet (h', r, t'). To learn such translation relation between entities, a margin-based ranking loss between the scores of correct and incorrect triplets $max(0, \gamma + f_r(h,t) - f_r(h',t'))$ is used in current translation based models, and has proven effective on knowledge graph embedding [7, 15, 16, 18, 33]. By the margin-based ranking loss, the score of a positive triplet

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'17 , November 6–10, 2017, Singapore, Singapore
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4918-5/17/11. . . $15.00
https://doi.org/10.1145/3132847.3132939

(h, r, t) is lower at least by γ than that of corresponding negative triplet, meanwhile a low score of correct triplet should be expected. However, we notice that, with the margin-based ranking loss it is also possible that the score of correct triplet is not small enough to hold (h, r, t), and even that $\mathbf{h} + \mathbf{r}$ maybe far away from $\mathbf{t}$. For example as given in subsection 3.1, when $\gamma - (f_r(h', t') - f_r(h, t))$ is very low, but $f_r(h, t)$ and $f_r(h', t')$ can be both much higher, which leads to the expectation $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ for golden triplet (h, r, t) cannot be realized. That is, the margin-based ranking loss favors a margin γ between the two scores $f_r(h, t)$ and $f_r(h', t')$, but cannot ensure the score $f_r(h, t)$ within an expected value domain. Using such loss the learned knowledge embeddings may lose the translation rule for correct fact (h, r, t). Common knowledge graph embedding methods mostly focus on improving the definition of score function, and ignore the issue of the margin-based ranking loss function to learn a suitable score of positive triplets.

In order to enhance the expectation of low scoring for positive triplets, this paper presents an upper limit score of positive triplets by a limit-based scoring loss, and adds the limit-based scoring loss item into common loss function as new loss evaluation for optimizations. In this way, two basic translation-based models, TransE [7] and TransH [33] are extended to TransE-RS and TransH-RS. The proposed TransE-RS and TransH-RS combine margin-based ranking loss and limit-based scoring loss, meanwhile separately share the same translation rules as TransE and TransH. The effectiveness of our proposed models are verified by our experiments in section 4.

Our contributions. (1) A limit-based scoring loss item is combined with margin-based ranking loss for translation-based models on learning knowledge embeddings. (2) This paper extends two simple and effective translation-based models TransE and TransH to TransE-RS and TransH-RS, which provides a reference for other translation-based models and other knowledge graph embedding models.

In the remainder of this paper, the related works on knowledge graph embedding are introduced in Section 2. We propose to improve translation-based models by introducing a limit-based scoring loss, and present TransE-RS and TransH-RS in Section 3. We detail the experimental studies on our models in Section 4, and finally give a conclusion in Section 5.

2 RELATED WORK

We introduce some typical models to learn knowledge embeddings in this section. All these methods embed entities into a vector space and enforce the embedding compatible under a relation r-dependent scoring. Different models differ in the definition of scoring functions $f_r(h, t)$ between h and t, but they have the same margin-based ranking loss framework.

2.1 Translation based Models

TransE [7] encodes entities and relations in the same space R^k, and regards the relation r as translation from h to t for a golden triplet (h, r, t). The score function of TransE, $f_r(h, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2$, is low if it is a golden triplet, and high otherwise. Such score is very efficient to 1-to-1 relation, but it has issues for N-to-1, 1-to-N and N-to-N relations. For example, by a 1-to-N relation, a head will only be translated to the same tail, that is, if r is a 1-to-N relation for $\{(h, r, t_i)\}_{i=1,2,\ldots,N}$, then $t_1 = t_2 = \ldots = t_N$, which does not comport with the facts.

TransH [33] introduces a mechanism of projecting entities into relation-specific hyperplane that enables different roles of an entity in different relations, to overcome the issue of TransE in modeling 1-to-N, N-to-1 and N-to-N relations. For a triplet (h, r, t), the projected entities $\mathbf{h}_\perp$ and $\mathbf{t}_\perp$ are connected by a translation vector with low error if (h, r, t) holds. The score function of TransH is defined as $f_r(h, t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$, where $\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^T \mathbf{h} \mathbf{w}_r$, $\mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^T \mathbf{t} \mathbf{w}_r$, and $\mathbf{h}, \mathbf{h}_\perp, \mathbf{t}, \mathbf{t}_\perp, \mathbf{r}, \mathbf{w}_r \in R^k$. $\mathbf{w}_r$ restricted with $\|\mathbf{w}_r\| = 1$ is the normal vector of the relation hyperplane. Although TransH extends modeling flexibility by employing relation hyperplanes, but similar to TransE it still assumes entities and relations within the same space.

TransR/CTransR [18] considers relations and entities as completely different objects in distinct semantic space, and its score function is $f_r(h, t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$, where $\mathbf{h}_r = \mathbf{h}\mathbf{M}_r$, $\mathbf{t}_r = \mathbf{t}\mathbf{M}_r$, $\mathbf{h}$ and $\mathbf{t}$ are in the entity space R^k, $\mathbf{h}_r$, $\mathbf{t}_r$, $\mathbf{r}$ are in r-relation subspace R^d, and $\mathbf{M}_r \in R^{d \times k}$ is the mapping matrix between the two spaces.

TransD [15] considers the multiple types of entities and relations simultaneously, and replaces transfer matrix by the product of two projection vectors of an entity-relation pair. The score function of TransD is $f_r(h, t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$ where $\mathbf{h}_\perp = \mathbf{M}_{rh}\mathbf{h}$, $\mathbf{t}_\perp = \mathbf{M}_{rt}\mathbf{t}$, $\mathbf{M}_{rh} = \mathbf{r}_p \mathbf{h}_p^T + \mathbf{I}^{m \times n}$, $\mathbf{M}_{rt} = \mathbf{r}_p \mathbf{t}_p^T + +\mathbf{I}^{m \times n}$.

TranSparse [16] adopts sparse matrices to model different types of relations, which considers the heterogeneity and the imbalance issues of knowledge graphs.

2.2 Other Models

Unstructured Model (UM) [5, 6] is a simplified case of TransE [7], which sets all translations r=0, i.e., the scoring function is $f_r(h, t) = \|\mathbf{h} - \mathbf{t}\|_2^2$. Obviously it cannot distinguish different relations.

Structured Embedding (SE) [4] introduces two relation-specific weight matrices for head and tail entities, i.e., $\mathbf{M}_{rh}$ and $\mathbf{M}_{rt}$. L1 distance between two projected vectors is defined as the score function, $f_r(h, t) = \|\mathbf{M}_{rh}\mathbf{h} - \mathbf{M}_{rt}\mathbf{t}\|_2^2$. This model is weak in capturing correlations between entities and relations as it uses two separate matrices.

Single Layer Model (SLM) [28] introduces nonlinear transformations by neural networks. It concatenates h and t as an input layer to a non-linear hidden layer then the linear output layer gives the resulting score: $f_r(h, t) = \mathbf{u}_r^T g(\mathbf{M}_{rh}\mathbf{h} + \mathbf{M}_{rt}\mathbf{t} + \mathbf{b}_r)$, where $\mathbf{M}_{rh}$ and $\mathbf{M}_{rt}$ are weight matrices, and $g(.)$ is the tanh operation. SLM is a special case of NTN when the tensor in NTN is set to 0.

Semantic Matching Energy (SME) [5, 6] aims to capture correlations between entities and relations via multiple matrix products and Hadamard product. SME considers two definitions of semantic matching energy functions for optimization, including the linear form $f_r(h,t) = (\mathbf{M}_1\mathbf{h}+\mathbf{M}_2\mathbf{r}+\mathbf{b}_1)^T(\mathbf{M}_3\mathbf{h}+\mathbf{M}_4\mathbf{r}+\mathbf{b}_2)$ and the bilinear form $f_r(h,t) = (\mathbf{M}_1\mathbf{h} \otimes \mathbf{M}_2\mathbf{r} + \mathbf{b}_1)^T(\mathbf{M}_3\mathbf{h} \otimes \mathbf{M}_4\mathbf{r} + \mathbf{b}_2)$. SME model shares the same parameters for all the relations, where $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$ and $\mathbf{M}_4$ are weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias vectors, $\otimes$ is the Hadamard product.

Latent Factor Model (LFM) [14, 29] considers second-order correlations between entity embeddings using a quadratic form, and defines a bilinear score function $f_r(h,t) = \mathbf{h}^T\mathbf{M}_r\mathbf{t}$.

NTN Model (NTN) [28] extends the Single Layer Model by considering the second-order correlations into nonlinear transformation (neural networks). The score function is $f_r(h,t) = \mathbf{u}_r^T g(\mathbf{h}^T\mathbf{M}_r\mathbf{t} + \mathbf{M}_{rh}\mathbf{h} + \mathbf{M}_{rt}\mathbf{t} + \mathbf{b}_r)$, where $\mathbf{u}_r$is a relation-specific linear layer, $g(\cdot)$ is the tanh operation, $\mathbf{M}_r \in R^{d\times d\times k}$is a 3-way tensor, and $\mathbf{M}_{rh}, \mathbf{M}_{rt} \in R^{k\times d}$ are weight matrices. However, the model complexity is much higher, making it difficult to handle large scale graphs.

3 OUR MODELS

Among the translation-based models mentioned above, TransE and TransH as basic models have low time complexities [15] and efficient predictive performance [7, 33]. In this paper, we present to extend TransE and TransH by combining limit-based scoring loss with margin-based ranking loss, separately to TransE-RS and TransH-RS. In this section, we firstly introduce the margin-based ranking loss and analyze its issues for learning knowledge embeddings.

3.1 Margin-based Ranking Loss

In translation-based models, to learn the entities embeddings and relations for fitting translation rules, a margin-based ranking criterion over the training set is defined [7, 15, 16, 18, 33]:

$$L_R = \sum_{(h,r,t)\in\Delta} \sum_{(h',r,t')\in\Delta'} [\gamma_1 + f_r(h,t) - f_r(h',t'))]_+ \quad (1)$$

where $[x]_+ = max(0,x)$ aims to get the maximum between 0 and x. Δ is the set of positive triplets, and $\Delta' = \{(h',r,t)|h' \in E\} \cup \{(h,r,t')|t' \in E\}$ denotes the set of corrupted triplets, which is composed of training triplets with either the head or tail replaced by a random entity. γ_1 is the margin separating positive and negative triplets.

The margin-based ranking loss function aims to make the score $f_r(h',t')$ of corrupted triplet higher by at least γ_1 than $f_r(h,t)$ of positive triplet. The parameters of the loss function will be updated when the loss is more than zero, otherwise not. For each training triplet (h,r,t) the loss function is try to realize

$$f_r(h',t') - f_r(h,t) \geq \gamma_1 \quad (2)$$

However, we note that such loss function maybe not ensure $f_r(h,t)$ to be low enough to represent the transaction between h and t. Factually from Eq.(2) it cannot be proved $f_r(h,t) < \varepsilon$ ($\varepsilon > 0$, ε is an arbitrarily small positive real number), i.e. $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$.

For example, under margin $\gamma_1 = 1$ a pair of scoring of positive and negative triplets $\{f_r(h',t'), f_r(h,t)\}$, can be $\{f_r(h',t') = 1.1, f_r(h,t) = 0.1\}$, $\{f_r(h',t') = 5.1, f_r(h,t) = 4.1\}$, and $\{f_r(h',t') = 30.1, f_r(h,t) = 29.1\}$. These pairs have the same zero loss, but in the last example the score of positive triplet (equal to 29.1) is much higher. Obviously a training triplet (h,r,t) cannot reach the golden condition $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$ by such high scoring of $f_r(h,t)$.

To overcome the issue of margin-based loss function for translation-based models, in the next section we present to supplement a scoring loss on the models to limit $f_r(h,t)$ of positive triplets.

3.2 Limit-based Scoring Loss

In order to effectively make the score of positive triplets within a low value limit, we propose to set the upper limit of the scoring for the correct triplet (h,r,t), by defining a limit-based scoring loss function:

$$L_S = \sum_{(h,r,t)\in\Delta} [f_r(h,t) - \gamma_2]_+ \quad (3)$$

The parameters of the scoring loss will be updated when the loss is more than zero, otherwise not. By such scoring loss function, $f_r(h,t)$ favors lower score than γ_2 for a correct triplet (h,r,t), i.e.

$$f_r(h,t) \leq \gamma_2 \quad (4)$$

3.3 TransE-RS and TransH-RS

This paper presents a new loss function framework, denoted as L_{RS} loss, which combines the limit-based scoring loss with the margin-based ranking loss as

$$L_{RS} = L_R + \lambda L_S, \quad (\lambda > 0) \quad (5)$$

for our extended translation based models, TransE-RS and TransH-RS.

TransE-RS and TransH-RS separately have the same definition of score function with TransE and TransH as follows:

(TransE-RS)

$$\begin{aligned} &f_r(h,t) = \|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2^2 \\ &\mathbf{h},\mathbf{t},\mathbf{r}, \in R^k \end{aligned} \quad (6)$$

(TransH-RS)

$$\begin{aligned} &f_r(h,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2 \\ &\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^T\mathbf{h}\mathbf{w}_r \\ &\mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^T\mathbf{t}\mathbf{w}_r \\ &\mathbf{h},\mathbf{h}_\perp,\mathbf{t},\mathbf{t}_\perp,\mathbf{r},\mathbf{w}_r \in R^k \end{aligned} \quad (7)$$

Optimization: The presented L_{RS} loss function Eq.(5) for TransE-RS and TransH-RS is rewritten as follows in detail:

Model	#Parameters	#Operations(Time complexity)
Unstructured(Bordes et al. 2012 [5])	$O(N_e m)$	$O(N_t)$
SE (Borders et al. 2011 [4])	$O(N_e m + 2N_r n^2)(m = n)$	$O(2m^2 N_t)$
SME (linear)(Borders et al. 2012 [5])	$O(N_e m + N_r n + 4mk + 4k)(m = n)$	$O(4mkN_t)$
SME(bilinear)(Borders et al. 2012 [5])	$O(N_e m + N_r n + 4mks + 4k)(m = n)$	$O(4mksN_t)$
LMF (Jenatton et al. 2012 [14])	$O(N_e m + N_r n^2)(m = n)$	$O((m^2 + m)N_t)$
SLM (Socher et al. 2013 [28])	$O(N_e m + N_r(2k + 2nk))(m = n)$	$O((2mk + k)N_t)$
NTN (Socher et al. 2013 [28])	$O(N_e m + N_r(n^2 s + 2ns + 2s))(m = n)$	$O(((m^2 + m)s + 2mk + k)N_t)$
TransE (Borders et al. 2013 [7])	$O(N_e m + N_r n)(m = n)$	$O(N_t)$
TransH (Wang et al. 2014 [33])	$O(N_e m + 2N_r n)(m = n)$	$O(2mN_t)$
TransR (Lin et al. 2015 [18])	$O(N_e m + N_r(m + 1)n)$	$O(2mnN_t)$
CTransR (Lin et al. 2015 [18])	$O(N_e m + N_r(m + d)n)$	$O(2mnN_t)$
TransD (Ji et al. 2015 [15])	$O(2N_e m + 2N_r n)$	$O(2nN_t)$
TransE-RS (this paper)	$O(N_e m + N_r n)(m = n)$	$O(N_t)$
TransH-RS (this paper)	$O(N_e m + 2N_r n)(m = n)$	$O(2mN_t)$

Table 1: Complexity (the number of parameters and the number of multiplication operations).

$$L_{RS} = \sum_{(h,r,t)\in\Delta} \sum_{(h',r,t')\in\Delta'} \{[\gamma_1 + f_r(h,t) - f_r(h',t')]_+ + \lambda[f_r(h,t) - \gamma_2]_+\} \tag{8}$$

The constraints of parameters for the former margin-based ranking loss in respective translation-based models are still preserved in the corresponding extended models. TransE-RS: $\forall \mathbf{e}$ in entity set, $\|\mathbf{e}\|_2 = 1$; $\forall \mathbf{r}$ in relation set, $\|\mathbf{r}\|_2 = 1$. TransH-RS: $\forall \mathbf{e}$ in entity set, $\|\mathbf{e}\|_2 \leq 1$; $\forall \mathbf{d_r}$, $|\mathbf{w}_r^T \mathbf{d}_r|/\|\mathbf{d}_r\|_2 \leq \epsilon$; $\forall \mathbf{w}_r$ for a relation, $\|\mathbf{w}_r\|_2 = 1$.

The optimization for minimizing the L_{RS} loss, with the constraints mentioned above, is carried out gradient descent over the possible entities, translation vectors and other parameters. When a golden triplet is visited, a negative triplet is randomly constructed according to the reference [33]. After a mini-batch, the gradient is computed and the model parameters are updated.

For a mini-batch of training triplets $\{(h_i, r_i, t_i)\}_{i=1\sim N_B}$, we can generate a general training set $\{(h_i, r_i, t_i), (h_i{}', r_i, t_i{}')\}$ by adding corrupted triplets. The loss of the mini-batch is

$$L_{RS} = \sum_i L_{RS}(i) \tag{9}$$

where

$$\begin{aligned} L_{RS}(i) =& L_R(i) + \lambda L_S(i) \\ =& [\gamma_1 + f_{ri}(h_i, t_i) - f_{ri}(h'_i, t'_i)]_+ \\ & + \lambda[f_{ri}(h_i, t_i) - \gamma_2]_+ \end{aligned} \tag{10}$$

The gradient of L_{RS} can be written as

$$\nabla L_{RS} = \sum_i \nabla L_{RS}(i) \tag{11}$$

where

$$\begin{aligned} \nabla L_{RS}(i) =& \nabla L_R(i) + \lambda \nabla L_S(i) \\ =& \nabla[\gamma_1 + f_{ri}(h_i, t_i) - f_{ri}(h'_i, t'_i)]_+ \\ & + \lambda \nabla[f_{ri}(h_i, t_i) - \gamma_2]_+ \end{aligned} \tag{12}$$

For a pair of positive and negative triplets $\{(h_i, r_i, t_i), (h_i{}', r_i, t_i{}')\}$, the $L_{RS}(i)$ loss includes two parts, margin-based ranking loss $L_R(i)$ and limit-based scoring loss $L_S(i)$. The gradient of $L_R(i)$ is also from the two loss, and the several cases of $\nabla L_{RS}(i)$ are given in follows:

Case 1 $\{(\gamma_1 + f_{r_i}(h_i, t_i) - f_{r_i}(h'_i, t'_i) > 0) \wedge (f_{r_i}(h_i, t_i) - \gamma_2 \leq 0)\}$: $\nabla L_{RS}(i) = \nabla[\gamma_1 + f_{r_i}(h_i, t_i) - f_{r_i}(h'_i, t'_i)] = \nabla f_{r_i}(h_i, t_i) - \nabla f_{r_i}(h'_i, t'_i)$;

Case 2 $\{ (\gamma_1 + f_{r_i}(h_i, t_i) - f_{r_i}(h'_i, t'_i) \leq 0) \wedge (f_{r_i}(h_i, t_i) - \gamma_2 > 0)\}$: $\nabla L_{RS}(i) = \lambda \nabla[f_{r_i}(h_i, t_i) - \gamma_2] = \lambda \nabla f_{r_i}(h_i, t_i)$;

Case 3 $\{ (\gamma_1 + f_{r_i}(h_i, t_i) - f_{r_i}(h'_i, t'_i) > 0) \wedge (f_{r_i}(h_i, t_i) - \gamma_2 > 0)\}$: $\nabla L_{RS}(i) = \nabla[\gamma_1 + f_{r_i}(h_i, t_i) - f_{r_i}(h'_i, t'_i)] + \lambda \nabla[f_{r_i}(h_i, t_i) - \gamma_2] = (1 + \lambda)\nabla f_{r_i}(h_i, t_i) - \nabla f_{r_i}(h'_i, t'_i)$;

Case 4 $\{otherwise\}$: $\nabla L_{RS}(i) = 0$.

Comparing to common L_R loss, our L_{RS} loss not only scores correct triplets lower by γ_1 than that corrupted triplets, but also scores correct triplets lower than γ_2.

For TransE-RS and TransH-RS, all embeddings for entities and relationships are first initialized following the random procedure proposed in [7, 10, 33], not using the results of TransE as [15, 18].

We also compare the parameter and operation complexities with several translation-based models and other baseline models reported by references [15, 16], shown in Table 1. N_e and N_r in this table are the number of entities and relations, N_t represents the number of triplets in a knowledge graph, m is the dimension of entity embedding spaces and n is the dimensions of relation embedding spaces. Same as TransE and TransH, TransE-RS and TransH-RS separately have low parameter complexities and time complexities.

4 EXPERIMENTS

We empirically study and evaluate our proposed models on two tasks, link prediction [7] and triplet classification [28] on subsection 4.1 and 4.2. The tasks are implemented on two popular knowledge graphs, WordNet [21] and Freebase [2]. WordNet is a large lexical knowledge graph, in which a synset as an entity expresses distinct concept and relationships between synsets indicate their lexical relations. Freebase is a large collaborative knowledge base consisting of a large number of world facts. Table 3 lists statistics of subsets from

Dataset	WN18				FB15k			
Metric	Mean		Hits@10 (%)		Mean		Hits@10(%)	
	raw	filt	raw	filt	raw	filt	raw	filt
Unstructured(Bordes et al. 2012 [5])	315	304	35.5	38.2	1074	979	4.5	6.3
RESCAL(Nickel, et al. 2011 [23])	1,180	1,163	37.2	52.8	828	683	28.4	44.1
SE (Borders et al. 2011 [4])	1,011	985	68.5	80.5	273	162	28.8	39.8
SME (linear)(Borders et al. 2012 [5])	545	533	65.1	74.1	274	154	30.7	40.8
SME(bilinear)(Borders et al. 2012 [5])	526	509	54.7	61.3	284	158	31.3	41.3
LMF (Jenatton et al. 2012 [14])	469	456	71.4	81.6	283	164	26.0	33.1
TransE (Borders et al. 2013 [7])	263	251	75.4	89.2	243	125	34.9	47.1
TransH (unif)(Wang et al. 2014 [33])	318	303	75.4	86.7	211	84	42.5	58.5
TransH (bern)(Wang et al. 2014 [33])	401	388	73.0	82.3	212	87	45.7	64.4
TransR (unif)(Lin et al. 2015 [18])	232	219	78.3	91.7	226	78	43.8	65.5
TransR (bern)(Lin et al. 2015 [18])	238	225	79.8	92.0	198	77	48.2	68.7
TransD (unif)(Ji et al. 2015 [15])	242	229	79.2	92.5	211	67	49.4	74.2
TransD (bern)(Ji et al. 2015 [15])	224	212	79.6	92.2	194	91	53.4	77.3
TransE-RS(unif)	362	348	**80.3**	**93.7**	**161**	**62**	**53.1**	**72.3**
TransE-RS(bern)	385	371	**80.4**	**93.7**	**161**	**63**	**53.2**	**72.1**
TransH-RS(unif)	401	389	**81.2**	**94.7**	**163**	**64**	**53.4**	**72.6**
TransH-RS(bern)	371	357	**80.3**	**94.5**	**178**	77	**53.6**	**75.0**

Table 2: Evaluation results on link prediction.

WordNet and Freebase in our experiments. WN11 [28] and WN18 [6] are two subsets of WordNet; and FB15k [6] and FB13 [28] are two subsets of Freebase.

Dataset	#Rel	#Ent	#Train	#Valid	#Test
WN11	11	38,696	112,581	2,609	10,544
WN18	18	40,943	141,442	5,000	5,000
FB13	13	75,043	316,232	5908	23,733
FB15k	1,345	14,951	483,142	50,000	59,071

Table 3: Datasets used in experiments.

We analyze the scoring distribution of positive and negative triplets by TransE, TransH and our extended models on Fb15K in subsection 4.3, and also discuss the effect of parameters γ_1, γ_2 and λ on our models in subsection 4.4.

4.1 Link prediction

Link prediction [5, 6] is to predict the missing h or t for a relation fact triple (h, r, t). In the experiments two datasets WN18 and FB15k (see Table 2) are used. For each testing triplet (h, r, t), we replace h or t entity by every entity in the knowledge graph and rank all the entities in descending order according to the scores calculated by score function. The correct entities for missing prediction should lead to lower $f_r(h, t)$ scores and meanwhile hit former ranks. The settings “raw” and “filt” distinguish whether or not to consider the impact of a corrupted triplet existing in correct KG. Following common translation based models, two metrics are reported: the averaged rank of correct entities (denoted as Mean), and the proportion of top-10 rank for correct entities (denoted as Hits@10). The expected results for a good model should be that “Mean” is lower and “Hits@10” is higher. For constructing the corrupted triples, we use “unif” to denote the traditional way of replacing head or tail with equal probability, and follow [33] to use “bern” to denote reducing false negative labels by replacing head or tail with different probabilities. For the compared models, as all the data sets are the same, we will refer to experimental results of several baselines from [7, 15, 18, 33].

In the experiments of our two proposal models, we select learning rate α for GD from {0.001, 0.005, 0.01}, parameter λ from {0.5, 1, 2}, ranking margin γ_1 from {0.25, 1, 2, 6}, upper limit γ_2 of scoring for positive triplet from $\{0.5\gamma_1, \gamma_1, 2\gamma_1, 3\gamma_1, 4\gamma_1\}$, the embedding dimension k from {50, 100}, the batch size B from {75, 120, 960, 1200, 4800}, L_1 distances for loss functions, and the weight C from {0.0625, 0.25, 1.0} for TransH-RS. The optimal parameters are determined by the validation set. We traverse all the training triplets for 1000 rounds.

WN18	α	λ	γ_1	γ_2	k	B
TransE-RS(“unif” & “bern”)	0.001	0.5	2	6	100	1200
TransH-RS(“unif” & “bern”)	0.001	1	2	6	100	1200

FB15K	α	λ	γ_1	γ_2	k	B
TransE-RS(“unif” & “bern”)	0.001	1	2	6	100	960
TransH-RS(“unif”)	0.001	1	2	6	100	960
TransH-RS(“bern”)	0.001	0.5	2	8	100	960

Table 4: Experimental Parameters for Link Prediction.

Evaluation results on both WN18 and FB15K are shown in Table 2. and the optimal configurations of TransE-RS and TransH-RS for link prediction are given in Table 4, and the weight C for TransH-RS are all set to 0.0625.

From Table 2, we can see that: (1) TransE-RS and TransH-RS outperform non-Translation models (Unstructured, SE, SME, and LMF), TransE, TransH and TransR, and also can be comparable to TransD except results of WN18_ Mean

	Predicting head(Hits@10)				Predicting tail(Hits@10)			
Relation Category	1-to-1	1-to-n	n-to-1	n-to-n	1-to-1	1-to-n	n-to-1	n-to-n
Unstructured(Bordes et al. 2012 [5])	34.5	2.5	6.1	6.6	34.3	4.2	1.9	6.6
SE (Borders et al. 2011 [4])	35.6	62.6	17.2	37.5	34.9	14.6	68.3	41.3
SME(linear)(Borders et al. 2012 [5])	35.1	53.7	19.0	40.3	32.7	14.9	61.6	43.3
SME(bilinear)(Borders et al. 2012 [5])	30.9	69.6	19.9	38.6	28.2	13.1	76.0	41.8
TransE(Borders et al. 2013 [7])	43.7	65.7	18.2	47.2	43.7	19.7	66.7	50.0
TransH (unif)(Wang et al. 2014 [33])	66.7	81.7	30.2	57.4	63.7	30.1	83.2	60.8
TransH (bern)(Wang et al. 2014 [33])	66.8	87.6	28.7	64.5	65.5	39.8	83.3	67.2
TransR (unif)(Lin et al. 2015 [18])	76.9	77.9	38.1	66.9	76.2	38.4	76.2	69.1
TransR (bern)(Lin et al. 2015 [18])	78.8	89.2	34.1	69.2	79.2	37.4	90.4	72.1
TransD (unif)(Ji et al. 2015 [15])	80.7	85.8	47.1	75.6	80.0	54.5	80.7	77.9
TransD (bern)(Ji et al. 2015 [15])	86.1	95.5	39.8	78.5	85.4	50.6	94.4	81.2
TransE-RS(unif)	**87.2**	**96.2**	35.9	**71.8**	**87.0**	45.0	**95.5**	**75.4**
TransE-RS(bern)	**87.4**	**96.3**	35.3	**71.7**	**86.5**	44.2	**95.4**	**75.2**
TransH-RS(unif)	**87.6**	**95.9**	35.6	**72.5**	**86.3**	44.9	**95.5**	**75.8**
TransH-RS(bern)	**85.6**	**95.5**	37.4	**75.5**	**85.7**	47.4	**94.9**	**78.7**

Table 5: Evaluation results on FB15K by mapping properties of relations.(%)

metric. (2) TransE-RS and TransH-RS perform better than other models on WN18_ Hit@10 metric, where our "raw" are higher than 80.3% and "filt" are higher than 93.7%. (3) TransE-RS and TransH-RS achieve great improvements on FB15k, the results of Hit@10("raw") are all more than 53.1% and that of Hit@10("filt") are all more than 72.1%. (4) We note that TransE-RS and TransH-RS have lower parameter complexities (see Table 1) than TransR and TransD.

For the comparison of Hits@10 of different kinds of relations, Table 5 shows the detailed results by mapping properties of relations following the same rules in [7] on FB15k. From Table 5, we can see that TransE-RS and TransH-RS obviously outperform TransE, TransH and TransR, and can be comparable to TransD. Especially TransE-RS and TransH-RS achieve higher accuracies on predicting head (including 1-to-1 and 1-to-n relations) and predicting tail (including 1-to-1 and n-to-1 relations). The accuracies of predicting head 1-to-1 are more than 85.6%, and that of 1-to-n are more than 95.5% . The accuracies of predicting tail 1-to-1 are more than 85.7%, and that of n-to-1 are more than 94.9%. Generally our proposed models obtain the highest accuracies among the compared methods on following items: predicting head (1-to-1 relations) "unif" is 87.6% and "bern" is 87.4%, predicting head (1-to-n relations) "unif" is 96.2% and "bern" is 96.3%, predicting tail (n-to-1 relations) "unif" is 87.0% and "bern" is 86.5%, and predicting tail (n-to-1 relations) "unif" is 95.5% and "bern" is 94.9%.

4.2 Triple Classification

This task aims to judge whether a given triple (h, r, t) is correct or not, i.e., binary classification on a triplet. It is used in [28] to evaluate NTN model on two datasets (WN11 and FB13) with small number of relations, and has been explored in [33] on FB15k containing much more relations. Following [33] , we also use three data sets, WN11, FB13 and FB15K (see Table 2), for the test of our models.

In the test phase, we need negative triples for the binary classification evaluation. The data sets WN11 and FB13 released by NTN [28] already have negative triples. For FB15k, we construct the negative triplets following the same procedure as in [28]. The decision rule for classification is that, a triplet (h,r,t) is predicted positive if the score f_r is below a relation-specific threshold, otherwise negative. The relation-specific threshold is optimized by maximizing classification accuracies on the validation set. For WN11 and FB13, we compare our models with baseline methods reported in [15, 18, 33] who used the same data sets. For FB15k, since our strategy for constructing negative triplets is same to [15, 18, 28], we did not rerun the compared baseline methods, and adopt their results reported in [15, 18]. For TransE-RS and TransH-RS, the initial entities are randomly given following [7, 10, 33].

For our proposal models we select learning rate α for GD from $\{0.001, 0.005, 0.01\}$, parameter λ from $\{0.5, 1, 2\}$, ranking margin γ_1 from $\{0.25, 1, 2, 6, 7\}$, upper limit γ_2 of scoring for positive triplet from $\{0.5\gamma_1, 0.7\gamma_1, \gamma_1, 2\gamma_1, 3\gamma_1, 4\gamma_1\}$, the embedding dimension k from $\{50, 100\}$, the batch size B from $\{120, 480, 960, 1200, 4800\}$, L_1 distances for loss functions, and the weight C from $\{0.0625, 0.25, 1.0\}$ for TransH-RS. The optimal parameters are determined by the validation set. We traverse all the training triplets for only 500 rounds, as there are no significant improvements with much more rounds.

Evaluation results of triple classification are shown in Table 6, and the optimal configurations of TransE-RS and TransH-RS for triplet classification are given in Table 7, and the weight C for TransH-RS are all set to 0.0625. From Table 6, TransE-RS and TransH-RS significantly outperform non-Translation models, TransE and TransH, and can be comparable to TransR and TransD for triplets classification. On WN11 our models with all possible settings can reach more than 85.2%, which improve TransE and TransH more

Data Sets	WN11	FB13	FB15K
SE	53.0	75.2	-
SME(bilinear)	70.0	63.7	-
SLM	69.9	85.3	-
LFM	73.8	84.3	-
NTN	70.4	87.1	68.2
TransE (unif)	75.9	70.9	77.3
TransE (bern)	75.9	81.5	79.8
TransH (unif)	77.7	76.5	74.2
TransH (bern)	78.8	83.3	79.9
TransR (unif)	85.5	74.7	81.1
TransR (bern)	85.9	82.5	82.1
TransD (unif)	85.6	85.9	86.4
TransD (bern)	86.4	89.1	88.0
TransE-RS(unif)	85.2	82.8	82.0
TransE-RS(bern)	85.3	83.0	81.9
TransH-RS(unif)	**86.3**	82.1	83.0
TransH-RS(bern)	**86.4**	81.6	83.2

Table 6: Experimental results on Triplets Classification Accuracies(%).

WN11	α	λ	γ_1	γ_2	k	B
TransE-RS("unif" & "bern")	0.01	1	7	4.9	100	120
TransH-RS("unif" & "bern")	0.01	1	10	7	100	120
FB13	α	λ	γ_1	γ_2	k	B
TransE-RS("unif" & "bern")	0.001	0.25	2	8	100	1200
TransH-RS("unif" & "bern")	0.001	1.5	2	8	100	1200
FB15K	α	λ	γ_1	γ_2	k	B
TransE-RS("unif" & "bern")	0.001	1	2	6	100	960
TransH-RS("unif")	0.001	1	2	6	100	960
TransH-RS("bern")	0.001	0.5	2	8	100	960

Table 7: Experimental Parameters for Link Prediction.

than 6.4%. On FB13 and FB15K, our models also have significant improvements compared to TransE, TransH and TransR, but cannot perform better than TransD in our selected optimal parameters range. Moreover, we should note that (1) TransE-RS and TransH-RS have same parameter and operation complexities as TransE and TransH, which is lower than TransR and TransD. (2) There are differences in initialing entities embeddings between our models and TransR/TransD. Our models randomly initial the entities, not use the learned embeddings by TransE as TransR and TransD. It means that our models have much better ability to overcome the problem of overfitting.

We also compared the classification accuracies of each relation by TransE, TransH, TransE-RS and TransH-RS on WN11. In this experiment, we rerun the TransE and TransH with the parameters reported in [33], obtain slightly different accuracies 76.5%(TransE) and 71.9%(TransH) with the reported results(in Table 5), and we ignore the differences derived from randomly experiments. The accuracies of eleven relations on WN11 are given separately in Figure 1.

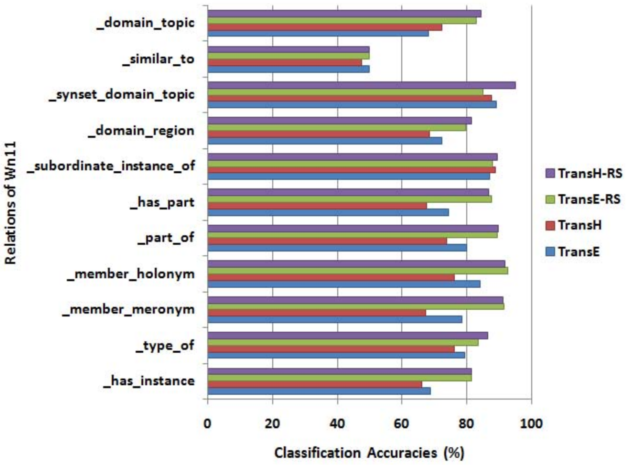


Figure 1: Classification accuracies of different relations on WN11

From results of Figure 1, TransE-RS and TransH-RS significantly improve TransE and TransH in each relation classification. TransH-RS is slightly better than TransE-RS.

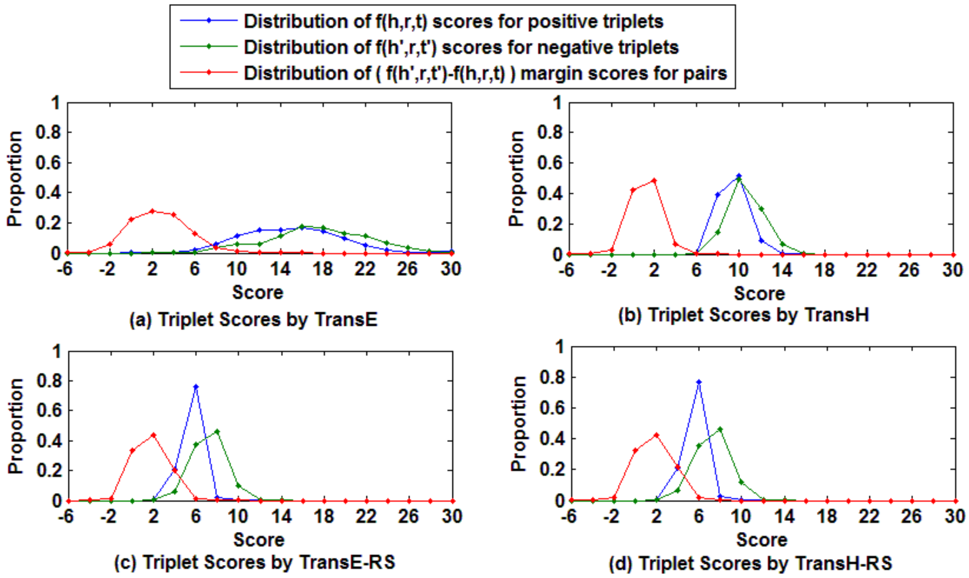
4.3 Distributions of Triplets' Scores

We discuss distributions of positive and negative triplets' scores by TransE, TransH, TransE-RS and TransH-RS, aiming to analyze the difference between L_R Loss and our L_{RS} Loss. On FB15K data set, we train knowledge embeddings with "unif" on training set, and use validation set to test the scoring distributions. There are 50000 correct triplets and corresponding 50000 corrupted triplets sampled by the procedure same to section 4.2 in [28]. For each pair of positive and negative triplets in validation set, we calculate the score $f_r(h,t)$ of positive triplet, the score $f_r(h',t')$ of negative triplet and the margin-score $f_r(h',t') - f_r(h,t)$ of the pair, and then give the distributions of three kinds of scores separately. In this experiment, during scoring range $[-6, 30]$, the scoring interval is set to 2, and for a score s, we count the proportion of triplets' scores in $(s-1, s+1]$ as the probability of score s. For example, for the distribution of positive triplets' score $f_r(h,t)$, the number of positive triplets during (s-1,s+1] is n_s, the $n_s/50000$ is as the proportion of the score s (where $s = -6, -4, -2, ., 30$) on $f_r(h,t)$ distribution.

The optimization parameters of TransE and TransH are same as reference [33], and the parameters of TransE-RS and TransH-RS on FB15K have been give in subsection 4.2. Table 8 only lists the parameters γ_1 for margin-scoring $f_r(h',t') - f_r(h,t)$ and γ_2 for scoring $f_r(h,t)$ of positive triplets in these models.

# Param	TransE	TransH	TransE-RS	TransH-RS
$\gamma 1$	1	0.25	2	2
$\gamma 2$	-	-	6	6

Table 8: Experimental Parameters.



(a) Triplet Scores by TransE

(b) Triplet Scores by TransH

(c) Triplet Scores by TransE-RS

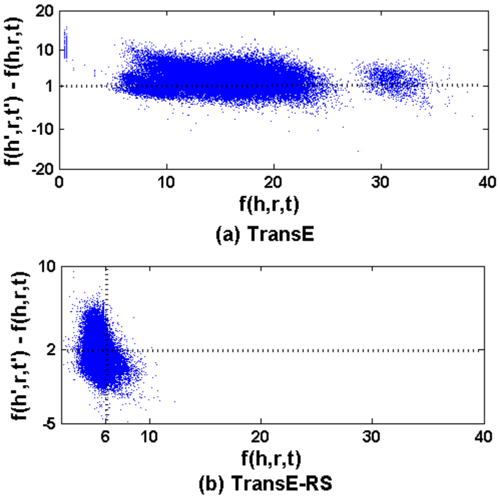
(d) Triplet Scores by TransH-RS

Figure 2: Distribution of triplets on different scores(FB15K)

Figure 2 (a), (b), (c) and (d) show separately the results of TransE, TransH, TransE-RS and TransH-RS on the distributions of three kinds of scores. In Figure 2 (c) and (d), TransE-RS and TransH-RS obviously have concentrated distributions on the scores of positive triplets and negative triplets. We give the analysis as follows: (1) "$\gamma_2 = 6$" used in L_S loss by our models means that the scoring $f_r(h,t)$ of positive triplets are expected lower than 6, thus there is high proportion score distribution of positive triplets smaller than 6 for TransE-RS and TransH-RS. About 80 percent of the positive triplets are scored lower than 6. (2) For TransE-RS and TransH-RS, the negative triplets also have high proportion distribution larger than the score 8, which is derived from the combination of L_R loss and L_S loss. The scoring of a positive triplet has high probability lower than 6, meanwhile the margin to the corresponding negative triplet is more than 2, so the negative triplet will have high probability larger than 8(6+2=8). See Figure 2 (c) and (d), about 46 percent of the negative triplets are scored larger than 8. (3) Moreover, the marginal scoring also have high proportion larger than the score 2, as the "$\gamma_1 = 2$" in L_R loss.

In Figure 2 (a) and (b), TransE only has a concentrated distribution for margin-score $f_r(h',t') - f_r(h,t)$, due to the margin-based ranking Loss L_R. But the other two distributions of scores $f_r(h,t)$ and $f_r(h',t')$ are not concentrated and rather scattered in scoring range from 6 to 30. Compared to TransE, TransH has the concentrated scorings separately on the $f_r(h,t)$ and $f_r(h',t')$ in Figure 2 (b). Compared to TransE and TransH, TransE-RS and TransH-RS have more centralized proportion distribution on the scoring of positive triplets, meanwhile maintain the margin between the scores of positive and negative triplets.

We further analyze the example of TransE and TransE-RS by another distribution map in Figure 3. We give score $f_r(h,t)$ and margin-score $f_r(h',t') - f_r(h,t)$ from a pair of positive and negative triplets, and then plot 2D point $(f_r(h,t), f_r(h',t') - f_r(h,t))$. For TransE, most points (about percent 70) $f_r(h',t') - f_r(h,t) > 1$, due to the parameter "$\gamma_1 = 1$" in L_R loss. For TransE-RS, there is not only "$\gamma_1 = 2$" in L_R loss, but also "$\gamma_2 = 6$" in L_S loss, thus percent 43 of points in the area of $(f_r(h',t') - f_r(h,t) > 2) \wedge (f_r(h,t) < 6)$,



(a) TransE

(b) TransE-RS

Figure 3: $\{f(h,r,t), (f(h',r,t') - f(h,r,t))\}$ distribution of triplets (FB15K)

and total percent of 89 points in the area of $f_r(h,t) < 6$. Compared with TransE, TransE-RS favors more concentrated and lower scores for positive triplets.

4.4 Discussion of Parameters

4.4.1 **Discussion on γ_1 and γ_2.** Different from common translation based models that only use margin-based ranking loss, our proposed models in addition consider another limit-based scoring loss. For the loss functions, γ_1 limits the score margin between positive and negative triplets, and γ_2 sets the upper limit of scoring for positive triplets. We also try to find the correlation of γ_1 and γ_2 from the results of link prediction and triplet classification. We find that $\gamma_2 = 3\gamma_1$ or $\gamma_2 = 4\gamma_1$ is better for link prediction, but for triplet classification there are not obvious characteristics on γ_1 and γ_2.

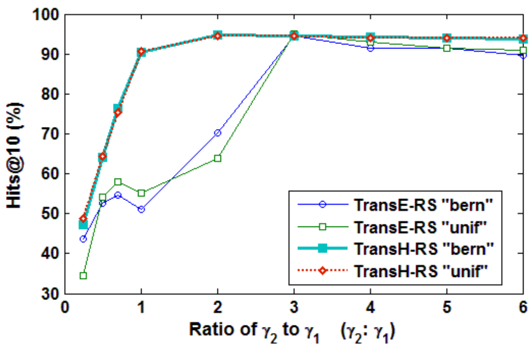
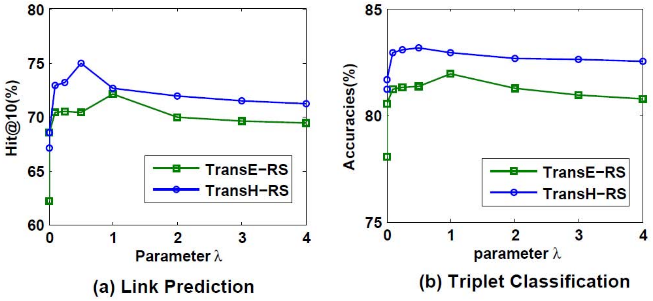


Figure 4: WN18 Hit10 Metric (”filt”) under different ratio of γ_2 to γ_1 with fixed γ_1.

In Figure 4, we give an example on WN18_Hit@10_ “filt” with different settings of γ_1 and γ_2. For TransE-RS and TransH-RS, $\gamma_1 = 2$, $\gamma_2 = \{0.25\gamma_1, 0.5\gamma_1, 0.75\gamma_1, \gamma_1, 2\gamma_1, 3\gamma_1, 4\gamma_1, 5\gamma_1, 6\gamma_1\}$, and other optimal configurations are $k = 100$, $B = 1200$, $\alpha = 0.001$, and $\lambda = 1$.

A lower γ_2 is expected to ensure the golden condition $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ for positive triplets, but we know that in complex knowledge graphs there exist many “1-to-n”, “n-to-1” and “n-to-n” relations, that is, an entity may be correlative to many relations. It means that an entity needs to satisfy many golden conditions at the same time, if the scoring of the entity’s positive triplet is limited to low value for one relation, it maybe lead to higher scoring for other relations. An expected γ_2 should make entities suit the evaluations from all the relations. Upon seeing Figure 4 we find that too lower γ_2 is not good for representation of knowledge embeddings, and also when $\gamma_2 = 3\gamma_1$ all the methods can reach a stable good results on Hit@10 (“filt”) metric. Moreover γ_1 maintains the discrimination between positive and negative triplets, which also give the rules on learning knowledge embeddings. Thus γ_1 and γ_2 are two important factors for our proposed models.

4.4.2 **Discussion on λ.** Parameter λ is used to combine L_S loss with L_R loss for our models, which is one of the different parameters from the former models. To analyze the influence of λ on our models, we test the two tasks link prediction and triplet classification under different λ on FB15K data set. We example “filt” and “bern” method in this experiment, λ is from $\{0, 0.001, 0.1, 0.25, 0.5, 1, 2, 3, 4\}$, and the other fixed parameters are same to our models on FB15K in subsection 4.1 and 4.2. We firstly learn models with different λ on training and validation set, and separately test the accuracies of link prediction and triplet classification on test set.



(a) Link Prediction

(b) Triplet Classification

Figure 5: Results with different λ on FB15K

Figure 5 (a) and (b) show the results of link prediction (Hit@10 metric) and triplet classification separately. From the results of Figure 5, we can see that different λ lead to various accuracies. For link prediction task, $\lambda = 1$ is best for TransE-RS (72.1%) and $\lambda = 0.5$ is best for TransH-RS (75.0%). For Triplet classification task, $\lambda = 1$ is best for TransE-RS (81.9%) and $\lambda = 0.5$ is best for TransH-RS (83.2%). From the general change of accuracies in Figure 5 (a) and (b), we find that the effect of λ on TransE-RS is greater than that on TransH-RS. That is, TransE-RS is more sensitive to parameter λ than TransH-RS. Thus TransH-RS is more stable than TransE-RS under λ settings.

5 CONCLUSIONS

For knowledge graph embedding, we propose a new loss framework which combines limit-based scoring loss and margin-based ranking loss to provide lower scoring of a golden triplet. By the new loss framework, we extend two basic translation-based models TransE and TransH, to TransE-RS and TransH-RS. Experimental results on triplet classification and link prediction show that the proposal TransE-RS and TrasH-RS significantly improve original translation-based models, and are comparable to state-of-the-art methods, meanwhile maintaining the low complexities of parameters. Our work can be used as a reference for other translation-based models and other knowledge graph embedding models.

6 ACKNOWLEDGMENTS

This work is supported by National Key R&D Program 2016 (No.2016YFB0801300), and the National Natural Science Foundation of China (No.61202226), We thank all anonymous reviewers for their constructive comments.

REFERENCES

[1] M. Ashburner, C. A. Ball, and J. A. et al. Blake. 2000. Geneontology: Tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *In Proceedings of KDD*. 1247–1250.

[3] A. Bordes, S. Chopra, and J. Weston. 2014b. Question answering with subgraph embeddings. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing,EMNLP*. 615–620.

[4] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. 2011. Learning structured embeddings of knowledge bases. In *In Proceedings of AAAI*. 301–306.

[5] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *In Proceedings of AISTATS*. 127–135.

[6] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. 2014a. A semantic matching energy function for learning with multirelational data. *Machine Learning* 94, 2 (2014a), 233–259.

[7] A. Bordes, N. Usunier, and A. Garcia-Duran. 2013. Translating Embeddings for Modeling Multi-relational Data. In *In Proceedings of NIPS*. 2787–2795.

[8] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

[9] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu. 2016. Neural Sentiment Classification with User and Product Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 1650–1659.

[10] X. Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. 249–253.

[11] S. Guo, Q. Wang, B. Wang, L. Wang, and L. Guo. 2015. Semantically Smooth Knowledge Graph Embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 84–94.

[12] S He, K. Liu, G. Ji, and J. Zhao. 2015. Learning to represent knowledge graphs with gaussian embedding. In *In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

[13] S. He, K. Liu, Y. Zhang, L. Xu, and J. Zhao. 2014. Question answering over linked data using first-order logic. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing,EMNLP*. 1092–1103.

[14] R. Jenatton, N. L. Roux, A. Bordes, and G. R. Obozinski. 2012. A latent factor model for highly multi-relational data. In *In Proceedings of NIPS*. 3167–3175.

[15] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *In Proceedings of ACL*. 687–696.

[16] G. Ji, K. Liu, S. He, and J. Zhao. 2016. Knowledge Graph Completion with Adaptive Sparse Transfer Matrix. In *In Proceedings of AAAI*. 985–991.

[17] Ni Lao, Tom M. Mitchell, and William W. Cohen. 2011. Random Walk Inference and Learning in A Large Scale Knowledge Base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. 529–539.

[18] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *In Proceedings of AAAI*. 2181–2187.

[19] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun. 2016. Neural Relation Extraction with Selective Attention over Instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 26 (2013), 3111–3119.

[21] G. A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM* 38, 11 (1995), 39–40.

[22] Maximilian Nickel, Volker Tresp, and Hans Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data. In *International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July*. 809–816.

[23] M. Nickel, V. Tresp, and H.-P. Kriegel. 2012. Factorizing yago: scalable machine learning for linked data. In *In Proceedings of WWW*. 271–280.

[24] Alberto Paccanaro and Geoffrey E Hinton. 2001. Learning Distributed Representations of Concepts Using Linear Relational Embedding. *IEEE Transactions on Knowledge and Data Engineering* 13 (2001).

[25] Hinrich Schuetze and Christian Scheible. 2013. Two SVDs produce more focal deep learning representations. *CoRR* abs/1301.3627 (2013).

[26] W. Shen, J. Wang, P. Luo, and M. Wang. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *In Proceedings of the 19th ACM SIGKDD*. 68–76.

[27] Farzaneh Shoeleh, Mahshid Majd, Ali Hamzeh, and Sattar Hashemi. 2015. Knowledge Representation in Learning Classifier Systems: A Review. *CoRR* abs/1506.04002 (2015).

[28] R. Socher, D. Chen, C. D. Manning, and A. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *In Proceedings of NIPS*. 926–934.

[29] I. Sutskever, J. B. Tenenbaum, and R. Salakhutdinov. 2009. Modelling relational data using bayesian clustered tensor factorization. In *In Proceedings of NIPS*. 1821–1828.

[30] Sean R. Szumlanski and Fernando Gomez. 2010. Automatically acquiring a semantic network of related concepts. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*. 19–28.

[31] Luke Vilnis and Andrew Mccallum. 2014. Word Representations via Gaussian Embedding. *Computer Science* (2014).

[32] William Yang Wang, Kathryn Mazaitis, Ni Lao, and William W. Cohen. 2015. Efficient inference and learning in a large knowledge base. *Machine Learning* (2015).

[33] Z. Wang, J. Zhang, J. Feng, and Z. Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *In Proceedings of AAAI*. 1112–1119.

[34] H. Xiao, M. Huang, and X. Zhu. 2016. From One Point to a Manifold: Knowledge Graph Embedding for Precise Link Prediction. In *IJCAI 2016,*. 1315–1321.

[35] H. Xiao, M. Huang, and X. Zhu. 2016. TransG: A Generative Model for Knowledge Graph Embedding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August 7-12, 2016, Berlin, Germany*. 2316C–2325.

[36] M. Yahya, K. Berberich, S. Elbassuoni, and G. Weikum. 2013. Robust question answering over the web of linked data. In *In Proceedings of the 22nd ACM international conference on CIKM*. Association for Computational Linguistics, 1107–1116.

[37] Limin Yao, Sebastian Riedel, and Andrew Mccallum. 2012. Probabilistic databases of universal schema. In *Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction*. 116–121.