Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Distributed Representation Learning for Knowledge Graphs with Entity Descriptions

Miao Fan, Qiang Zhou, Thomas Fang Zheng and Ralph Grishman

ELSEVIER

Recent studies of knowledge representation attempt to project both entities and relations, which originally compose a high-dimensional and sparse knowledge graph, into a continuous low-dimensional space. One canonical approach *TransE* (Bordes et al., 2013) which represents entities and relations with vectors (embeddings), achieves leading performances solely with triplets, i.e. (*head_entity, relation, tail_entity*), in a knowledge base. The cutting-edge method *DKRL* (Xie et al., 2016) extends *TransE* via a enhancing the embeddings with entity descriptions by means of deep neural network models. However, *DKRL* requires extra space to store parameters of inner layers, and relies on more hyper-parameters to be tuned. Therefore, we create a single-layer model which requests much fewer parameters. The model measures the probability of each triplet along with corresponding entity descriptions, and learns contextual embeddings of entities, relations and words in descriptions simultaneously, via maximizing the loglikelihood of the observed knowledge. We evaluate our model in the tasks of knowledge graph completion and entity type classification with two benchmark datasets: **FB500K** and **EN15K**, respectively. Experimental results demonstrate that the proposed model outperforms both *TransE* and *DKRL*, indicating that it is both efficient and effective in learning better distributed representations for knowledge bases.

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- A probabilistic model learns contextual embedding for knowledge graphs with text.
- Semantic relatedness among entities, relations and words is captured.
- A single-layer neural network model requires fewer parameters.
- It acquires better embeddings which assist the task of knowledge graph completion.
- It produces better features for the task of entity type classification.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Distributed Representation Learning for Knowledge Graphs with Entity Descriptions

Miao Fan[a,**], Qiang Zhou[a], Thomas Fang Zheng[a], Ralph Grishman[b]

[a] *CSLT, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, China.*
[b] *Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, NY, 10003, U.S.A.*

ABSTRACT

Recent studies of knowledge representation attempt to project both entities and relations, which originally compose a high-dimensional and sparse knowledge graph, into a continuous low-dimensional space. One canonical approach *TransE* (Bordes et al., 2013) which represents entities and relations with vectors (embeddings), achieves leading performances solely with triplets, i.e. (*head_entity, relation, tail_entity*), in a knowledge base. The cutting-edge method *DKRL* (Xie et al., 2016) extends *TransE* via enhancing the embeddings with entity descriptions by means of deep neural network models. However, *DKRL* requires extra space to store parameters of inner layers, and relies on more hyperparameters to be tuned. Therefore, we create a single-layer model which requests much fewer parameters. The model measures the probability of each triplet along with corresponding entity descriptions, and learns contextual embeddings of entities, relations and words in descriptions simultaneously, via maximizing the loglikelihood of the observed knowledge. We evaluate our model in the tasks of knowledge graph completion and entity type classification with two benchmark datasets: **FB500K** and **EN15K**, respectively. Experimental results demonstrate that the proposed model outperforms both *TransE* and *DKRL*, indicating that it is both efficient and effective in learning better distributed representations for knowledge bases.

1. Introduction

A typical large-scale knowledge base, such as Freebase (Bollacker et al., 2008), mainly contains billions of triplets (*head_entity, relation, tail_entity*), abbreviated as (*h, r, t*), and each represents a fact that there is a relation *r* between the two entities (*h* and *t*). These triplets conventionally compose a knowledge graph in which each entity is a node, and relations between two entities are regarded as directed edges. This symbolic representation facilitates storing and displaying knowledge, but makes the inference of knowledge infeasible, especially when the volume of knowledge base grows and data becomes sparse.

Therefore, recent research on knowledge representations attempts to address the issue via projecting both entities and relations into a continuous low-dimensional space (García-Durán et al., 2016; Nickel et al., 2016). One canonical approach is *TransE* (Bordes et al., 2013), which solely uses triplets in a knowledge base without requiring extra text to make the inference of knowledge computable. It learns low-dimensional vector representations (embeddings) of both entities and relations by minimizing a margin-based loss function. *TransE* attracts many successive studies (Fan et al., 2014, 2015c,d; Lin et al., 2015; Wang et al., 2014b), not only because of its leading performances, but also owing to fewer parameters required.

However, knowledge is expected to reinforce other intelligent applications, such as question-answering (QA) systems (Shekarpour et al., 2016) in which unstructured text is also involved. On the other hand, mainstream knowledge repositories, such as Freebase (Bollacker et al., 2008) and NELL (Carlson et al., 2010), contain concise entity descriptions and relation mentions in addition. The extra text provides contextual evidence to help learn better embeddings. Therefore, the study of context-enhanced representation learning for knowledge bases becomes prosperous (Fan et al., 2015a,b; Wang et al., 2014a; Weston et al., 2013). The cutting-edge method *DKRL* (Xie et al., 2016) extends *TransE* via enhancing the embeddings with

** Corresponding author: Tel.: +86-135-817-00448;
e-mail: fanmiao.cslt.thu@gmail.com (Miao Fan)

entity descriptions by means of deep neural network models. However, *DKRL* demands extra space to store parameters of inner layers, and relies on more hyperparameters to be tuned.

In this paper, we create a single-layer model which requires much fewer parameters for representation learning of knowledge bases (*RLKB*). *RLKB* measures the probability of each triplet along with corresponding entity descriptions. During the phase of training, the model learns contextual embeddings of entities, relations and words in descriptions simultaneously via maximizing the loglikelihood of the whole observed knowledge base, so as to encoding those embeddings into the same low-dimensional vector space. We evaluate our model with two benchmark datasets: **FB500K** and **EN15K**. **FB500K** contains nearly 500,000 triplets from Freebase, and it is a wide-spread dataset adopted by many recent studies (Bordes et al., 2013; Fan et al., 2014, 2015c,d; Lin et al., 2015; Wang et al., 2014b) to test the performance of knowledge graph completion. **EN15K** is composed by almost 15,000 entities occurring in **FB500K** for the task of entity type classification (Xie et al., 2016), and each entity usually belongs to multiple principal entity types in Freebase. Experimental results demonstrate that the embeddings acquired by *RLKB* outperforms both *TransE* and *DKRL* in the tasks of knowledge graph completion and entity type classification, indicating that the proposed model is both efficient and effective in learning better knowledge representations. We also explore the reason why *RLKB* achieves such leap forwards on the two experimental tasks, and find out that the acquired embeddings can obtain semantic relatedness among entities, relations, and even words in descriptions. Intuitively, the semantic relatedness revealed by *RLKB* within knowledge repositories, not only helps narrow down the scope of searching missing entities/relations, but also provides similar representations between entities that share the same types.

Overall, we contribute a single layer neural network model which requires fewer parameters to learn knowledge embeddings. In order to acquire contextual embeddings, both structured knowledge graphs and unstructured text are used, so that the semantic relatedness among entities, relations and even words is captured. These embeddings learnt by our approach not only assist the task of knowledge graph completion, but also produce better features for the task of entity type classification.

2. Related Work

The trend of studying distributed representations for knowledge bases has emerged in recent years. A group of articles (Bordes et al., 2013; Fan et al., 2014, 2015c,d; Lin et al., 2015; Wang et al., 2014b) engage in exploring representation models based on structured knowledge graphs without requiring extra text, and we will talk about them in Section 2.1. Section 2.2 is going to give a review on context-enhanced approaches, in which entity descriptions (Wang et al., 2014a; Xie et al., 2016) and relation mentions (Fan et al., 2015a,b; Weston et al., 2013) are additionally considered while learning knowledge embeddings.

2.1. Structure-based Representation Learning

TransE (Bordes et al., 2013) is a classical approach on learning vector representations of both entities and relations solely with knowledge graphs. The approach regards relations between two entities as translating operations in vector spaces, and uses the scoring function $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ to measure the plausibility of each triplet (h, r, t). Its strength lies in requiring fewer parameters to represent triplets. However, *TransE* cannot cope well with multi-relations between two entities, since these relations tend to gain the same embeddings. To address the issue, Wang et al. propose *TransH* (Wang et al., 2014b) in which two entities are projected into different relation-dependent hyperplanes, so that each relation can distinguish from the others. Fan et al. (Fan et al., 2014) simply adapt the learning rates along with the number of multi-relations, and achieve great improvements. Several state-of-the-art models, such as *IIKE* (Fan et al., 2015c), *LMNNE* (Fan et al., 2015d) and *TransR* (Lin et al., 2015), are created to learn better embeddings of entities and relations within knowledge graphs, but none of them considers extra information from text, such as entity descriptions and relation mentions which are included in most knowledge repositories as well.

2.2. Context-enhanced Representation Learning

Weston et al. (Weston et al., 2013) firstly concern about encoding words in relation mentions together with entities and relations. They match the mentions with corresponding relations, based on the assumption of distant supervision (Mintz et al., 2009). In addition, Fan et al. (Fan et al., 2015a,b, 2016) leverage the relation mentions already aligned by NELL, and propose several jointly embedding models.

On the other hand, Wang et al. (Wang et al., 2014a) start to align entities with anchors in Wikipedia to obtain contextual descriptions. This framework has constraints of application scenario, because linking by entity names severely pollutes the embeddings of words, and using Wikipedia anchors completely relies on the special data source. Therefore, Xie et al. (Xie et al., 2016) propose *DKRL* which directly uses the concise descriptions of entities in Freebase, and achieves state-of-the-art performances on the tasks of knowledge graph completion and entity type classification. However, *DKRL* adopts deep neural network models to refine embeddings of entities and descriptions, so that it needs more space to store parameters of inner layers and more hyperparameters to be tuned. Therefore, we design a single-layer model which will be described in the subsequent section to address those issues.

3. Model

Given a knowledge repository Δ which contains enormous number of items (h, r, t, d_h, d_t), where each item is composed by a head entity h with its descriptions d_h, a tail entity t with its descriptions d_t, and a relation r between the two entities, our model aims to maximize the loglikelihood of the observed Δ, expressed by Eq. (1), to obtain the embeddings of entities, relations and words in entity descriptions:

$$\arg\max_{h,r,t,d_h,d_t} \sum_{(h,r,t,d_h,d_t)\in\Delta} \log Pr(h, r, t, d_h, d_t), \tag{1}$$

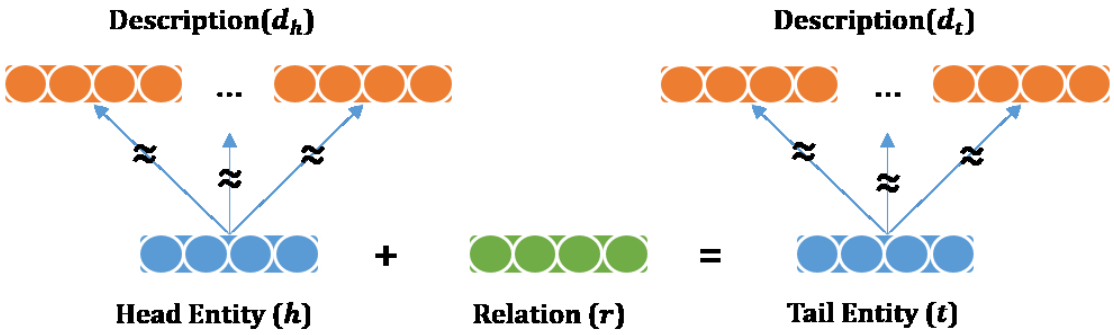

Fig. 1. The framework of *RLKB*, a single layer model in which the embeddings of entities, relations, and keywords in descriptions are jointly trained.

in which the probability of each item $(h, r, t, d_h, d_t) \in \Delta$ is influenced by two factors:

$$\log Pr(h, r, t, d_h, d_t) = \log Pr(h, r, t) + \log Pr(d_h, d_t|h, r, t). \quad (2)$$

$Pr(h, r, t)$ represents the probability of the observed triplet (h, r, t), and $Pr(d_h, d_t|h, r, t)$ is the conditional probability of observing entity descriptions given the triplet.

If a triplet (h, r, t) is a positive example, we need to make sure that any one of the three objects (h, r and t) is plausible given the other two. In other words, $Pr(h, r, t)$ is the trade-off among the three conditional probabilities:

$$\log Pr(h, r, t) = \frac{\log Pr(h|r, t) + \log Pr(r|h, t) + \log Pr(t|h, r)}{3}. \quad (3)$$

To represent $Pr(h|r, t)$, the conditional probability of observing h given r and t, we firstly adopt the geometric modeling of a triplet proposed by *TransE* (Bordes et al., 2013) which is illustrated by Fig. 1: The relation ($\mathbf{r}$) between two entities is considered as a translation in the vector space from the head entity ($\mathbf{h}$) to the tail entity ($\mathbf{t}$). Furthermore, we define a scoring function Θ as follow:

$$\Theta(h, r, t) = \alpha - \frac{1}{2}\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2 \quad (4)$$

to quantify the plausibility of a triplet, and α is a positive bias for structured knowledge. The larger Θ scores, the higher possibility that the evaluated triplet is positive. It is obvious that many incorrect/negative head entities, denoted by h', are likely to replace h and to collapse the triplet. Suppose that the set of these negative head entities is E'_h. We define $Pr(h|r, t)$ as

$$Pr(h|r, t) = \frac{\exp^{\Theta(h,r,t)}}{\exp^{\Theta(h,r,t)} + \sum_{h' \in E'_h} \exp^{\Theta(h',r,t)}}. \quad (5)$$

We can also use the identical way to define $Pr(r|h, t)$ and $Pr(t|h, r)$ as follows,

$$Pr(r|h, t) = \frac{\exp^{\Theta(h,r,t)}}{\exp^{\Theta(h,r,t)} + \sum_{r' \in R'} \exp^{\Theta(h,r',t)}}, \quad (6)$$

and

$$Pr(t|h, r) = \frac{\exp^{\Theta(h,r,t)}}{\exp^{\Theta(h,r,t)} + \sum_{t' \in E'_t} \exp^{\Theta(h,r,t')}}, \quad (7)$$

in which r' is a negative relation included by the set of negative relations R', and t' is a negative tail entity belonging to the set of negative tail entities E'_t.

However, it is difficult to calculate the normalizers of $\log Pr(h|r, t)$, $\log Pr(r|h, t)$ and $\log Pr(t|h, r)$, since the number of items included in E'_h, R', and E'_t usually reach millions. Fortunately, we can use the negative sampling method instead, proposed by Mikolov et al. (Mikolov et al., 2013; Goldberg and Levy, 2014), to approximate $\log Pr(h|r, t)$:

$$\log Pr(h|r, t) \approx \log Pr(1|h, r, t) + \frac{1}{n}\sum_{i=1}^{n} \log Pr(0|h'_i, r, t), \quad (8)$$

in which n represents the number of negative head entities randomly picked up from E'_h. Follow the Eq. (8), we also transform $\log Pr(r|h, t)$ and $\log Pr(t|h, r)$:

$$\log Pr(r|h, t) \approx \log Pr(1|h, r, t) + \frac{1}{n}\sum_{i=1}^{n} \log Pr(0|h, r'_i, t), \quad (9)$$

$$\log Pr(t|h, r) \approx \log Pr(1|h, r, t) + \frac{1}{n}\sum_{i=1}^{n} \log Pr(0|h, r, t'_i). \quad (10)$$

Here $Pr(1|h, r, t)$ indicates the probability of the assertion "the triplet (h, r, t) is positive" is true:

$$Pr(1|h, r, t) = \frac{1}{1 + \exp^{-\Theta(h,r,t)}}, \quad (11)$$

and

$$Pr(0|h'_i, r, t) = \frac{1}{1 + \exp^{\Theta(h'_i,r,t)}} \quad (12)$$

indicates the possibility that the triplet (h'_i, r, t) is corrupted.

Besides dealing with the triplets (h, r, t), we need to model $Pr(d_h, d_t|h, r, t)$ as well. We assume that descriptions (d_h or d_t) are composed according to the target entities (h or t):

$$\log Pr(d_h, d_t|h, r, t) = \log Pr(d_h|h) + \log Pr(d_t|t). \quad (13)$$

Furthermore, we extract m keywords by means of TF-IDF, which is short for term frequency-inverse document frequency (Manning et al., 2008), to well represent descriptions. Suppose that $d_h = \{w_1, w_2, ..., w_m\}$ (e.g., in which w_2 denotes the second extracted keyword), we try to make the embedding of the head

entity close to the embeddings of its keywords as shown by Fig. 1 as well. We define another function $\Phi(d_h|h)$ to measure the distance:

$$\Phi(d_h|h) = \beta - \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{h} - \mathbf{w}_j\|_2^2, \tag{14}$$

in which β is another positive bias for unstructured text. Then we define $Pr(d_h|h)$ as:

$$Pr(d_h|h) = \frac{\exp^{\Phi(d_h|h)}}{\exp^{\Phi(d_h|h)} + \sum_{d'_h \in D'_h} \exp^{\Phi(d'_h|h)}}, \tag{15}$$

in which D'_h is the set of negative descriptions d'_h that do not belong to h. According to the theory of negative sampling (Mikolov et al., 2013; Goldberg and Levy, 2014), $\log Pr(d_h|h)$ is eventually transformed into:

$$\log Pr(d_h|h) \approx \log Pr(1|d_h, h) + \frac{1}{n}\sum_{i=1}^{n} \log Pr(0|d'_{h,i}, h), \tag{16}$$

where

$$Pr(1|d_h, h) = \frac{1}{1 + \exp^{-\Phi(d_h,h)}}. \tag{17}$$

And $\log Pr(d_t|t)$ can be approximated in the same way:

$$\log Pr(d_t|t) \approx \log Pr(1|d_t, t) + \frac{1}{n}\sum_{i=1}^{n} \log Pr(0|d'_{t,i}, t), \tag{18}$$

in which

$$Pr(1|d_t, t) = \frac{1}{1 + \exp^{-\Phi(d_t,t)}}. \tag{19}$$

To sum up, Eq. (20) shows the decomposed formula which approximates the joint loglikelihood of an item (h, r, t, d_h, d_t).

$$\begin{aligned}
&\log Pr(h, r, t, d_h, d_t) \\
&\approx \log Pr(1|h, r, t) + \log Pr(1|d_h, h) + \log Pr(1|d_t, t) \\
&+ \frac{1}{3n}\sum_{i=1}^{n}\{\log Pr(0|h'_i, r, t) + \log Pr(0|h, r'_i, t) + \log Pr(0|h, r, t'_i)\} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\{\log Pr(0|d'_{h,i}, h) + \log Pr(0|d'_{t,i}, t)\}.
\end{aligned} \tag{20}$$

4. Algorithm

We use Eq. (20) to replace the original objective function which is shown by Eq. (1), and obtain the subsequent object to maximize the loglikelihood of all observed items in a knowledge repository Δ:

$$\begin{aligned}
&\arg\max_{h,r,t,d_h,d_t} \sum_{(h,r,t,d_h,d_t)\in\Delta} \{\log Pr(1|h, r, t) + \log Pr(1|d_h, h) + \log Pr(1|d_t, t) \\
&+ \frac{1}{3n} \sum_{(h',r',t',d'_h,d'_t)\in\Delta'_{(h,r,t,d_h,d_t)}} [\log Pr(0|h', r, t) + \log Pr(0|h, r', t) \\
&+ \log Pr(0|h, r, t')] \\
&+ \frac{1}{n} \sum_{(h',r',t',d'_h,d'_t)\in\Delta'_{(h,r,t,d_h,d_t)}} [\log Pr(0|d'_h, h) + \log Pr(0|d'_t, t)]\},
\end{aligned} \tag{21}$$

in which (h', r', t', d'_h, d'_t) is a negative item sampled from the set of n negative samples $\Delta'_{(h,r,t,d_h,d_t)}$, given a positive item (h, r, t, d_h, d_t).

The new objective function makes it much easier to obtain the partial derivatives of all knowledge embeddings, i.e. $\mathbf{h}, \mathbf{r}, \mathbf{t}, \mathbf{d_h}$ and $\mathbf{d_t}$. We use Stochastic Gradient Ascent (SGA) algorithm to achieve the object of *RLKB*, and to update better distributed representations in the meanwhile. Algorithm 1 displays the steps of learning the *RLKB* model written in pseudocode.

Specifically, the embeddings of entities, relations and words are firstly initialized obeying the same uniform distribution (Step 1-9). Then each positive triplet along with its entity descriptions is selected from the training set, and we randomly generate n negative triplets with descriptions based on the positive sample (Step 12-14). To update the embeddings of entities, relations and words, we further conduct gradient ascent algorithm according to Eq. (21). We do the steps from 11 to 17 in iterative fashion until we find the embeddings of entities, relations and words in descriptions that produce the highest probability over the validation set (Step 17).

5. Experiments

To evaluate the quality of the embeddings acquired by knowledge representation learning, researchers generally compare the performances of using them to support several application scenarios, such as completing the missing entities or relations in a knowledge graph, and providing better features for predicting types of entities.

Besides the tasks of knowledge graph completion and entity type classification which are used to demonstrate the effectiveness of *RLKB*, we compare the complexity of all the models involved.

5.1. Knowledge Graph Completion

Knowledge graph completion is a classical task which aims at completing a triplet (h, r, t) when one of the three items is missing. *TransE* (Bordes et al., 2013) mainly focuses on inferencing head or tail entities. *DKRL* (Xie et al., 2016) extends the scope to relation prediction which attempts to recover the missing relation(s) between two observed entities.

5.1.1. Dataset

The dataset **FB500K** inherits most triplets from the benchmark dataset **FB15K** constructed by *TransE*. However, **FB15K** only contains structured triplets extracted from Freebase. To bring in entity descriptions, *DKRL* maps each entity in **FB15K** to Freebase, and filters a few entities without adequate descriptions. We rename the new dataset **FB500K** after its volume, and extract Top 5 keywords for each entity as "descriptions" by means of TF-IDF method. Table 1 shows the statistics of the new benchmark dataset **FB500K** in which each entity has a five-keyword description.

Algorithm 1 : The Learning Algorithm of *RLKB*

Input:

The training knowledge base $\Delta = \{(h, r, t, d_h, d_t)\}$, entity set E, relation set R, vocabulary set V of entity descriptions; dimension of embeddings k, number of negative samples n, learning rate r, the bias α and β.

foreach $\mathbf{e} \in E$ **do**
  $\mathbf{e} := \text{Uniform}(-\frac{6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
foreach $\mathbf{r} \in R$ **do**
  $\mathbf{r} := \text{Uniform}(-\frac{6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
foreach $\mathbf{w} \in V$ **do**
  $\mathbf{w} := \text{Uniform}(-\frac{-6.0}{\sqrt{k}}, \frac{6.0}{\sqrt{k}})$
end foreach
while not adequate rounds **do**
  foreach $(h, r, t, d_h, d_t) \in \Delta$ **do**
    foreach $i \in \text{range}(n)$ **do**
      $\Delta'_{(h,r,t,d_h,d_t)}$ appends a negative sample: $\langle h'_i, r'_i, t'_i, d'_{h,i}, d'_{t,i} \rangle$
      /*$\Delta'_{(h,r,t,d_h,d_t)}$ is the set of n negative samples, given the positive knowledge (h, r, t, d_h, d_t).*/
    end foreach
    Conduct gradient ascent with learning rate r on $\log Pr(h, r, t, d_h, d_t)$, and update the embeddings based on Eq. (20).
  end foreach
  Check the probability over the validation set.
  Set $\Delta'_{(h,r,t,d_h,d_t)}$ empty.
end while

Output:

All the embeddings of h, t, r, w, where $h, t \in E$, $r \in R$ and $w \in \{d_h, d_t\}$.

Table 1. Statistics of the benchmark dataset: FB500K for knowledge graph completion.

#Train	#Valid	#Test	#Entity	#Relation	#Vocabulary
472,860	48,991	57,803	14,904	1,341	27,144

Table 2. Experimental results of entity inference on FB500K.

APPROACH	MEAN RANK		MEAN HIT@10(%)	
	Raw	Filter	Raw	Filter
TransE (Bordes et al., 2013)	232	141	35.6	43.0
DKRL(CBOW) (Xie et al., 2016)	236	151	38.3	51.8
DKRL(CNN) (Xie et al., 2016)	200	113	44.3	57.6
RLKB	**160**	**52**	**49.7**	**64.6**

Table 3. Experimental results of relation prediction on FB500K.

APPROACH	MEAN RANK		MEAN HIT@1(%)	
	Raw	Filter	Raw	Filter
TransE (Bordes et al., 2013)	3.52	3.25	60.6	71.6
DKRL(CBOW) (Xie et al., 2016)	2.85	2.51	65.3	82.7
DKRL(CNN) (Xie et al., 2016)	2.91	2.55	**69.8**	**89.0**
RLKB	**2.78**	**2.44**	65.9	83.3

5.1.2. Evaluation protocols

Following the protocols proposed by *TransE* and *DKRL* to measure the performances of entity inference and relation prediction, we take turns to replace the head entity, the relation and the tail entity of each testing triplet, with all the other entities and relations that appear in the training set. Then, we obtain three bunches of candidate triplets for each testing triplet. Within each bunch, we compute the scores of candidate triplets using the function $\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$ to measure its plausibility as knowledge, and then sort them in ascending order. Finally, we locate the ground-truth triplet and record its rank. For the task of entity inference, we concern on two metrics, i.e. *Mean Rank* and *Mean Hit@10* (the proportion of ground truth triplets that rank in Top 10), to measure the performance. Since there are much fewer relations in the two datasets, we use *Mean Hit@1* instead of *Mean Hit@10* in the task of relation prediction. However, the results measured by those *raw* metrics are relatively inaccurate, as the procedures above tend to generate false negative triplets. In other words, some of the candidate triplets rank rather higher than the ground truth triplet just because they also appear in the training set. We thus *filter* out those triplets to report more reasonable results as well.

5.1.3. Hyperparameter settings

Several hyperparameters need to be tuned by the validation set. They are: the dimension k of embeddings; the learning rate r of Stochastic Gradient Ascend (SGA) algorithm; number of negative samples n; the bias α in Eq. (4) and β in Eq. (14). We select our model with $k \in \{50, 100, 150, 200\}$, $r \in \{0.01, 0.025, 0.5, 1.0\}$ $n \in \{5, 10, 15\}$, $\alpha \in \{6.0, 8.0, 10.0\}$ and $\beta \in \{6.0, 8.0, 10.0\}$, and finally find that $k = 150$, $r = 0.025$, $n = 15$ and $\alpha = \beta = 8.0$ is the combination making the model perform the best on the validation set.

5.1.4. Results

We copy the best experimental results of state-of-the-art method *DKRL* (Xie et al., 2016), in which the authors conduct experiments with two deep neural networks, i.e. CBOW and CNN, on the **FB500K** dataset. For the baseline approach *TransE*, we run the original code[1] posted by Bordes et al. (Bordes et al., 2013) and set the best hyperparameters reported by them. Table 2 shows the results of entity inference task performed by the baseline, state-of-the-art approaches and the proposed *RLKB*. We observe that *RLKB* achieves a great leap forward on the metrics of *Mean Rank* and *Mean Hit@10*. Specifically, our model relatively improves 54.0% on *Filter Mean Rank* and 12.2% Top-10 ground-truth triplets completed by entities. The improvements indicate that *RLKB* can learn better embeddings of entities which are closer to their desired positions in the vector space.

We also use the same embeddings acquired from *TransE*, *DKRL*, and *RLKB* with the best hyperparameter settings to conduct relation prediction task. Table 3 demonstrates that *RLKB* outperforms the other approaches on the metric of *Mean Rank*, but shows comparable results with *DKRL(CBOW)* on the metric of *Hit@1*. Overall, our model relatively improves 4.3% on *Filter Mean Rank*, but decreases 6.4% Top-10 ground-truth triplets completed by relations.

Table 4. Statistics of the benchmark dataset: EN15K for entity type classification.

#Train	#Test	#Entity Type
12,113	1,332	50

5.2. Entity Type Classification

Entity Type Classification is a recently proposed task for evaluating the quality of knowledge embeddings. It is originally designed by (Neelakantan and Chang, 2015) and further adopted by *DKRL* (Xie et al., 2016). This task uses the learnt distributed representations as features to train a multi-label classifier, since most entities belong to multiple types.

5.2.1. Dataset

We use the same dataset built by *DKRL* (Xie et al., 2016) and rename it **EN15K** after its volume. **EN15K** is constructed upon the entities that appear in **FB500K**. Firstly, all entity types within **FB500K** are extracted from Freebase, and we collect 4,054 types in total. Then we pick up the Top 50 types by their frequency, and remove the *common*/*topic* which every entity has. Finally, we gain 13,445 entities covered by the most frequent 50 types. As Table 4 shows, Xie et al. (Xie et al., 2016) split these entities into the training set and the test set, which has 12,113 entities and 1,332 entities, respectively.

5.2.2. Evaluation protocols

Both (Neelakantan and Chang, 2015) and (Xie et al., 2016) use Mean Average Precision (*MAP*)[2] (Manning et al., 2008) as the standard metric to evaluate the performance of entity type classification. It is a common evaluation protocol adopted by multi-label classification problems.

5.2.3. Hyperparameter settings

We keep the hyperparameter settings of knowledge graph completion task, and directly take advantages of the best embeddings acquired by *RLKB*. The embeddings of entities are fed as features into an one-versus-rest Logistic Regression classifier implemented by Scikit-learn (Pedregosa et al., 2011). To gain the best performance, we further use five-fold cross validation with the training set to tune the hyperparameters, including the inverse of regularization strength $C \in \{0.1, 1.0, 10.0\}$, and the norm used in the penalization *penalty* $\in \{'l1', 'l2'\}$. Finally, we achieve the best combination of hyperparameters for *RLKB*: $C = 0.1$ and *penalty* $= 'l2'$.

5.2.4. Results

Table 5 shows the results of entity type classification evaluated by the metric of *MAP* with the **EN15K** dataset. We take turn-

[1] https://github.com/glorotxa/SME

[2] http://web.stanford.edu/class/cs276/handouts/EvaluationNew-handout-6-per.pdf

Table 5. Experimental results of entity type classification on EN15K.

APPROACH	MAP
BOW	0.863
TransE (Bordes et al., 2013)	0.882
DKRL(CBOW) (Xie et al., 2016)	0.893
DKRL(CNN) (Xie et al., 2016)	0.901
RLKB	**0.922**

s to feed BOW (Bag-of-words) features and embeddings learnt by *TransE*, *DKRL* and *RLKB* into the one-versus-rest Logistic Regression classifier, and observe that *RLKB* outperforms the state-of-the-art approach *DKRL*. Our model increases 2.1% on the metric of Mean Average Precision, indicating that the embeddings of entities generated by *RLKB* tend to close to each other, if they share the same types.

5.3. Model Complexity Comparison

The experimental results in Section 5.1 and Section 5.2 demonstrate that *RLKB* is more effective than the state-of-the-art *DKRL* and the baseline approach *TransE*, when conducting tasks of knowledge graph completion and entity type classification. In this part, we want to further analyze the cost of achieving the leading performances, via comparing the model complexities among the approaches we have mentioned.

Table 6. Comparison of model complexity among *TransE*, *DKRL(CBOW)*, *DKRL(CNN)*, and *RLKB*.

APPROACH	#Parameters
TransE (Bordes et al., 2013)	$O(n_e k + n_r k)$
DKRL(CBOW) (Xie et al., 2016)	$O(n_e k + n_r k + n_w k)$
DKRL(CNN) (Xie et al., 2016)	$O(n_e k + n_r k + n_w k + lk^2)$
RLKB	$O(n_e k + n_r k + n_w k)$

Table 6 shows that *TransE* needs the least space to store parameters, as it only focuses on encoding the entities and relations in a knowledge graph. Therefore, the model complexity of *TransE* is solely influenced by the number of entities n_e, the number of relations n_r, and the identical dimension k of vector representations. As *DKRL(CBOW)* and *RLKB* consider the words in entity descriptions in addition, their model complexities increase along with the number of words n_w in entity descriptions. Though they require almost the same number of parameters, but *RLKB* can achieve much better knowledge embeddings than *DKRL(CBOW)* according to the experimental results above.

Because *DKRL(CNN)* adopts an l-layer convolutional neural network to learn the embeddings of entity descriptions, it has to maintain extra space for the parameters of inner layers apart from the essential memories to store the embeddings of entities, relations, and words.

6. Discussions

Besides observing the quantitative results performed by the two tasks, i.e. knowledge graph completion and entity type classification, we look forward to exploring the essence of learning distributed representations of knowledge bases. Our model assumes itself capable of encoding not only structured knowledge graph information, but also unstructured text descriptions, into continuous vector spaces, so that we can bridge the gap of one-hot representations, and expect to discover certain relevance among entities, relations and even keywords in descriptions.

An intuitive way of revealing the relevance is to measure the L_2-norm distance between embeddings. For example, if we search the Top 10 nearest entities to $/m/01n4w\_$ (Washington and Lee University) which is a private liberal arts university in Lexington, Virginia, United States, we can gain a ranked list of universities shown by Table 7, instead of any other entity types. We can also find out that these nearest entities, to some extent, capture semantic similarities with $/m/01n4w\_$ from different aspects: For instance, they either share the same words in names, or locate at the same places.

This phenomenon also exists among relations, and words in descriptions, as proved by Table 8 and Table 9, respectively. The discovery helps to explain the reason why our model can lead a huge leap forward compared with baseline and state-of-the-art approaches, when it conducts the tasks of knowledge graph completion and entity type classification. To sum up, 1) we can encode entities, relations and words in the same low-dimensional vector space with the help of the proposed model *RLKB*, and obtain semantic relatedness among the embeddings; The semantic relatedness 2) not only helps narrow down the scope of searching missing entities/relations, but also 3) provides similar representations between entities with the same types.

7. Conclusion and Future Work

This paper contributes an efficient and effective single-layer model for learning distributed representations (embeddings) of entities, relations and even words within entity descriptions. The proposed model acquires better low-dimensional vector representations via maximizing the loglikelihood of all triplets and corresponding entity descriptions in a knowledge base. Experimental results show that the embeddings learnt by our model help achieve better performances than the state-of-the-art and baseline approaches, when we use them to conduct tasks of knowledge graph completion and entity type classification with two benchmark datasets, respectively. We further explore the essence of learning distributed representations of knowledge bases, and find out that the embeddings can capture the semantic relatedness among entities, relations, and even words in descriptions. This discovery also explains the reason why the embeddings acquired by our model lead a leap forward in the tasks of knowledge graph completion and entity type classification.

Several open questions that we cannot answer with *RLKB* are expected to explore in the future, such as how to discriminate the embeddings of multi-relations between two entities (Wang et al., 2014b), or how to model the knowledge described by a sequence of relations (García-Durán et al., 2015).

Table 7. Top10 nearest entities to $/m/01n4w\_$ (Washington and Lee University) searched by the L_2-norm distance between embeddings.

ENTITY	$/m/01n4w\_$ (Washington and Lee University)
NEAREST@10	$/m/0kw4j$ (American University)
	$/m/017v3q$ (College of William & Mary)
	$/m/01nnsv$ (George Washington University)
	$/m/0pspl$ (Georgetown University)
	$/m/0438f$ (James Madison University)
	$/m/037s9x$ (Washington & Jefferson College)
	$/m/02zr0z$ (Virginia Union University)
	$/m/0g8rj$ (University of Virginia)
	$/m/07t90$ (University of Washington)
	$/m/04wlz2$ (Hampton University)

Table 8. Top10 nearest relations to $/award/award\_winner/awards\_won./award/award\_honor/award\_winner$ searched by the L_2-norm distance between embeddings.

RELATION	$/award/award\_winner/awards\_won./award/award\_honor/award\_winner$
NEAREST@10	$/music/performance\_role/track\_performances./music/track\_contribution/role$
	$/base/popstra/celebrity/dated./base/popstra/dated/participant$
	$/base/popstra/celebrity/friendship./base/popstra/friendship/participant$
	$/music/performance\_role/regular\_performances./music/group\_membership/role$
	$/base/popstra/celebrity/canoodled./base/popstra/canoodled/participant$
	$/people/person/spouse\_s./people/marriage/spouse$
	$/location/location/adjoin\_s./location/adjoining\_relationship/adjoins$
	$/award/award\_nominee/award_nominations./award/award\_nomination/award\_nominee$
	$/award/award\_nominated\_work/award\_nominations./award/award\_nomination/nominated\_for$
	$/influence/influence\_node/influenced$

Table 9. Top10 nearest keywords to *colledges* searched by the L_2-norm distance between embeddings.

KEYWORD	*colleges*
NEAREST@10	*laboratory*
	universities
	university
	instituion
	nonsectarian
	granting
	eiga
	students
	drumlins

Acknowledgements

This work is supported by China Scholarship Council, National Program on Key Basic Research Project (973 Program) under Grant 2013CB329304, National Science Foundation of China (NSFC) under Grant No.61433018 and No.61373075, Proteus Project of NYU. The first author conducted this research while he was a joint-supervision Ph.D. student of Tsinghua University and New York University.

References

Bollacker, K.D., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge, in: Wang, J.T. (Ed.), Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008, ACM. pp. 1247–1250. URL: http://doi.acm.org/10.1145/1376616.1376746, doi:10.1145/1376616.1376746.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data, in: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pp. 2787–2795.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M., 2010. Toward an architecture for never-ending language learning, in: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010).

Fan, M., Cao, K., He, Y., Grishman, R., 2015a. Jointly embedding relations and mentions for knowledge population, in: Angelova, G., Bontcheva, K., Mitkov, R. (Eds.), Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria, RANLP 2015 Organising Committee / ACL. pp. 186–191. URL: http://aclweb.org/anthology/R/R15/R15-1026.pdf.

Fan, M., Zhou, Q., Abel, A., Zheng, T.F., Grishman, R., 2015b. Probabilistic belief embedding for knowledge base completion. CoRR abs/1505.02433. URL: http://arxiv.org/abs/1505.02433.

Fan, M., Zhou, Q., Abel, A., Zheng, T.F., Grishman, R., 2016. Probabilistic belief embedding for large-scale knowledge population. Cognitive Computation .

Fan, M., Zhou, Q., Chang, E., Zheng, T.F., 2014. Transition-based knowledge graph embedding with relational mapping properties, in: Aroonmanakun, W., Boonkwan, P., Supnithi, T. (Eds.), Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 28, Cape Panwa Hotel, Phuket, Thailand, December 12-14, 2014, The PACLIC 28 Organizing Committee and PACLIC Steering Committee / ACL / Department of Linguistics, Faculty of Arts, Chulalongkorn University. pp. 328–337. URL: http://aclweb.org/anthology/Y/Y14/Y14-1039.pdf.

Fan, M., Zhou, Q., Zheng, T.F., 2015c. Learning embedding representations for knowledge inference on imperfect and incomplete repositories. CoRR abs/1503.08155. URL: http://arxiv.org/abs/1503.08155.

Fan, M., Zhou, Q., Zheng, T.F., Grishman, R., 2015d. Large margin nearest neighbor embedding for knowledge representation, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, 2015 - Volume I, IEEE. pp. 53–59. URL: http://dx.doi.org/10.1109/WI-IAT.2015.125, doi:10.1109/WI-IAT.2015.125.

García-Durán, A., Bordes, A., Usunier, N., 2015. Composing relationships with translations, in: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics. pp. 286–290. URL: http://aclweb.org/anthology/D/D15/D15-1034.pdf.

García-Durán, A., Bordes, A., Usunier, N., Grandvalet, Y., 2016. Combining two and three-way embedding models for link prediction in knowledge bases. J. Artif. Intell. Res. (JAIR) 55, 715–742. URL: http://dx.doi.org/10.1613/jair.5013, doi:10.1613/jair.5013.

Goldberg, Y., Levy, O., 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. CoRR abs/1402.3722. URL: http://arxiv.org/abs/1402.3722.

Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X., 2015. Learning entity and relation embeddings for knowledge graph completion, in: Bonet, B., Koenig, S. (Eds.), Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., AAAI Press. pp. 2181–2187. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571.

Manning, C.D., Raghavan, P., Schütze, H., et al., 2008. Introduction to information retrieval. volume 1. Cambridge university press Cambridge.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. CoRR abs/1301.3781. URL: http://arxiv.org/abs/1301.3781.

Mintz, M., Bills, S., Snow, R., Jurafsky, D., 2009. Distant supervision for relation extraction without labeled data, in: Su, K., Su, J., Wiebe, J. (Eds.), ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, The Association for Computer Linguistics. pp. 1003–1011. URL: http://www.aclweb.org/anthology/P09-1113.

Neelakantan, A., Chang, M., 2015. Inferring missing entity type instances for knowledge base completion: New dataset and methods, in: Mihalcea, R., Chai, J.Y., Sarkar, A. (Eds.), NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, The Association for Computational Linguistics. pp. 515–525. URL: http://aclweb.org/anthology/N/N15/N15-1054.pdf.

Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E., 2016. A review of relational machine learning for knowledge graphs. Proceedings of the IEEE 104, 11–33. URL: http://dx.doi.org/10.1109/JPROC.2015.2483592, doi:10.1109/JPROC.2015.2483592.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Shekarpour, S., Lukovnikov, D., Kumar, A.J., Endris, K., Singh, K., Thakkar, H., Lange, C., 2016. Question answering on linked data: Challenges and future directions. arXiv preprint arXiv:1601.03541 .

Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014a. Knowledge graph and text jointly embedding, in: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL. pp. 1591–1601. URL: http://aclweb.org/anthology/D/D14/D14-1167.pdf.

Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014b. Knowledge graph embedding by translating on hyperplanes, in: Brodley, C.E., Stone, P. (Eds.), Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada., AAAI Press. pp. 1112–1119. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531.

Weston, J., Bordes, A., Yakhnenko, O., Usunier, N., 2013. Connecting language and knowledge bases with embedding models for relation extraction, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL. pp. 1366–1371. URL: http://aclweb.org/anthology/D/D13/D13-1136.pdf.

Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M., 2016. Representation learning of knowledge graphs with entity descriptions, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12C17, 2016, Phoenix, Arizona USA.