IEICE TRANS. INF. & SYST., VOL.E100–D, NO.7 JULY 2017

PAPER

Zero-Shot Embedding for Unseen Entities in Knowledge Graph

Yu ZHAO[†*], ***Student Member***, **Sheng GAO**[†a)], **Patrick GALLINARI**[††], ***and*** **Jun GUO**[†], ***Nonmembers***

SUMMARY Knowledge graph (KG) embedding aims at learning the latent semantic representations for entities and relations. However, most existing approaches can only be applied to KG completion, so cannot identify relations including unseen entities (or Out-of-KG entities). In this paper, motivated by the zero-shot learning, we propose a novel model, namely JointE, jointly learning KG and entity descriptions embedding, to extend KG by adding new relations with Out-of-KG entities. The JointE model is evaluated on entity prediction for zero-shot embedding. Empirical comparisons on benchmark datasets show that the proposed JointE model outperforms state-of-the-art approaches. The source code of JointE is available at `https://github.com/yzur/JointE`.

key words: *zero-shot learning, knowledge graph, embedding learning, relation prediction*

1. Introduction

The knowledge graph (KG) is a special kind of structured databases for knowledge management, such as WordNet [1], Freebase [2]. They consist of a huge amount of knowledge triples in the form of (**subject entity, predicate relation, object entity**), or the abbreviation (s, p, o). Figure 1 provides an example for a knowledge fact (*Ithaca College, /location/location/containedby, New York*). A well studied problem is how to accomplish KG completion. Most existing approaches [3]–[9] are only able to estimate the relation between existing entities in the KG by learning the embeddings for all observed entities and predicate relations, that is, the embeddings for s, p, o. However, they are unable to estimate the relation including Out-of-KG entities which are unseen in KG, since they only learn the representations of In-KG entities which already have existed in KG. Knowledge graph extension (KGE) by predicting additional triple in the zero-shot scenario, in which at least one of entities is Out-of-KG entity, is a very challenging problem, due to lack of Out-of-KG entities' embeddings.

It can be summarized to the scenario of the zero-shot learning [10]. Zero-shot learning refers to the generic problem how to predict unseen labels. (It's an extreme case of transfer learning.) For example, in the image classification task, although the class "cat" does not exist in the training data set, can we still tell if an image in the testing data is a cat or not? It sounds impossible at the first glance, but it is possible by utilizing description information for prediction. For example, given the description that a cat has four legs and pointy ears, the learner might be able to make a correct prediction on a test image if it is a cat, without having seen a cat before [11].

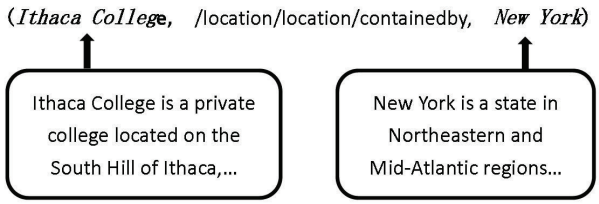

Fig. 1 Example of knowledge fact and the entity descriptions from Freebase.

In the spirit of zero-shot learning, the key motivation in this paper to solve the KGE problem is to utilize the descriptions for entities, that are available for most KGs. For example, in Fig. 1, subject entity (*Ithaca College*) and object entity (*New York*) have their descriptions in the box respectively. The descriptions usually explain entities, including rich semantic information about the entities. In this paper, we propose a novel model, namely JointE, jointly learning KG and entity descriptions embeddings, to extend KG by adding new relations with Out-of-KG entities. The main contributions in this paper are highlighted as follows:

- A novel model (JointE) is proposed to jointly learn the latent semantic representations for all entities and relations from KG and entity descriptions, which is able to estimate the relations involving unseen entities.
- Empirical comparison validates the effectiveness of the proposed model.

2. Related Works

KG embedding. Several energy-based models [3]–[9] have been proposed recently to encode the entities and relations of the triples into latent embedding space, i.e. KG embedding, for KG completion. These models are showed in Table 1. In order to clarify the difference between our proposed method and the existing methods, we also present the model of our proposed method JointE for KG embedding in Table 1. In addition, Zhang et al. [12], [13] propose models for

Manuscript received November 3, 2016.
Manuscript revised March 3, 2017.
Manuscript publicized April 10, 2017.
[†]The authors are with Beijing University of Posts and Telecommunications, Beijing 100876, China.
[††]The author is with LIP6, Universit Pierre et Marie Curie, Paris, France.
[*]Presently, with University of Rochester, USA.
a) E-mail: gaosheng@bupt.edu.cn (Corresponding author)
DOI: 10.1587/transinf.2016EDP7446

Table 1 Models for KG embedding.

Models	Scoring function $\mathbf{G}(\mathtt{s},\mathtt{p},\mathtt{o})$	Paremeters
SE [4]	$\|\mathbf{W}_{p1}\mathbf{s}-\mathbf{W}_{p2}\mathbf{o}\|_1$	$\mathbf{W}_{p1},\mathbf{W}_{p2}\in\mathbb{R}^{\kappa\times\kappa}$, $\mathbf{s},\mathbf{o}\in\mathbb{R}^{\kappa}$
SME [3]	$(\mathbf{W}_1\mathbf{s}\otimes\mathbf{W}_{p,1}\mathbf{p}+\mathbf{b}_1)^\top\cdot(\mathbf{W}_2\mathbf{o}\otimes\mathbf{W}_{p,2}\mathbf{p}+\mathbf{b}_2)$	$\mathbf{s},\mathbf{p},\mathbf{o}\in\mathbb{R}^{\kappa}$, $\mathbf{b}_1,\mathbf{b}_2\in\mathbb{R}^{\kappa}$ $\mathbf{W}_1,\mathbf{W}_{p,1},\mathbf{W}_2,\mathbf{W}_{p,2}\in\mathbb{R}^{\kappa\times\kappa}$
NTN [5]	$\mathbf{u}_p^\top\mathbf{f}(\mathbf{s}^\top\mathbf{W}_p^{[1:l]}\mathbf{o}+\mathbf{V}_p\begin{bmatrix}\mathbf{s}\\\mathbf{o}\end{bmatrix}+\mathbf{b}_p)$	$\mathbf{V}_p\in\mathbb{R}^{l\times 2\kappa}$ and $\mathbf{u}_p\in\mathbb{R}^{l}$, $\mathbf{b}_p\in\mathbb{R}^{l}$, $\mathbf{f}=tanh$
TransE [6]	$\|\mathbf{s}+\mathbf{p}-\mathbf{o}\|_d$	$\mathbf{s},\mathbf{p},\mathbf{o}\in\mathbb{R}^{\kappa}$
TransH [7]	$\|(\mathbf{s}-\mathbf{w}_p^\top\mathbf{s}\mathbf{w}_p)+\mathbf{p}-(\mathbf{o}-\mathbf{w}_p^\top\mathbf{o}\mathbf{w}_p)\|_d$	$\mathbf{s},\mathbf{o},\mathbf{w}_p,\mathbf{p}\in\mathbb{R}^{\kappa}$
TransR [8]	$\|\mathbf{s}_p+\mathbf{p}-\mathbf{o}_p\|_d$	$\mathbf{M}_\mathbf{p}\in\mathbb{R}^{k\times l}$, $\mathbf{s},\mathbf{o}\in\mathbb{R}^{k}$, $\mathbf{p}\in\mathbb{R}^{l}$ $\mathbf{s}_p=\mathbf{s}\mathbf{M}_\mathbf{p},\mathbf{o}_p=\mathbf{o}\mathbf{M}_\mathbf{p}$
Our proposed	$\mathbf{s}_d\cdot\mathbf{p}_s+\mathbf{p}_o\cdot\mathbf{o}_d+\mathbf{s}_d\cdot\mathbf{o}_d$	$\mathbf{s}_d,\mathbf{o}_d,\mathbf{p}_s$ and $\mathbf{p}_o\in\mathbb{R}^{\kappa}$

KG and text jointly embedding.

Word embedding. In Natural Language Processing (NLP), many language models [14]–[22] have been proposed for learning semantic knowledge from a huge amount of free text corpus, as encoding each word (phrase or sentence) to a semantic vector representation, namely word embedding. Word embedding can be used for various NLP tasks [16] such as POS tagging, chunking, named entity recognition, semantic and syntactic similarity [20], [21].

Zero-shot learning. [23] proposed a model DKRL, combining the existing model TransE (originally used for KG completion) [3] and CNN (or BOW), for KGE in zero-shot scenario. In computer vision, [24], [25] train a recognition model for zero-shot object recognition by specifying the category's attributes. [26] proposes a label-embedding model for attribute-based zero-shot classification. In recommendation systems, it is very challenging to recommend items to new users with no buying/rating history. Such zero-shot learning problem is called as cold start problem. Some existing approaches [27]–[29] are proposed to solve it.

3. Approach

We first present problem formulation of KGE in the zero-shot scenario. Next, we propose a new approach to build zero-shot embeddings of entities according to their descriptions. Finally, we propose a novel model, namely JointE, jointly learning embedddings from KG and entity descriptions.

3.1 KGE in the Zero-Shot Scenario

In this paper, we aim at KGE in the zero-shot scenario by predicting object entity given subject entity and predicate relation pair as `(subject, predicate, ?)` or predicting subject entity given predicate relation and object entity pair as `(?, predicate, object)`. In the zero-shot scenario, at least one entity in predicted triple is Out-of-KG entity. We propose a model (JointE) to jointly learning embeddings from knowledge facts and entity descriptions. By learning JointE model, we obtain the embeddings of In-KG entities, predicate relations and words, and then we build the representations for Out-of-KG entities according to their descriptions. If all the entities' representations have already learnt, we get a list of predicted Top-N objects for the test triple `(subject s, predicate p,?)` as follows:

$$\mathbf{Top}(\mathtt{s},\mathtt{p},\mathtt{N}):=\arg\max_{\mathtt{o}\in\hat{\mathcal{E}}}{}^{\mathtt{N}}\ \mathbf{G}(\mathtt{s},\mathtt{p},\mathtt{o})\,,$$

where N is the number of predicted entities, $\hat{\mathcal{E}}$ is used to denote sets all entities (including In-KG entities and Out-of-KG entities). $\mathcal{R}$ is used to denote the set of predicate relations. We use $\mathtt{s},\mathtt{o}$ and $\mathtt{p}$ to denote subject entity, object entity, and predicate relation respectively. $\mathbf{G}(\mathtt{s},\mathtt{p},\mathtt{o}):\hat{\mathcal{E}}\times\mathcal{R}\times\hat{\mathcal{E}}\rightarrow\mathbb{R}$ is the scoring function. Similarly, we obtain $\mathbf{Top}(\mathtt{N},\mathtt{p},\mathtt{o})$ for test triple `(?, predicate p, object o)`. We show the detailed methods of the model in the rest of this section.

3.2 Zero-Shot Embedding from Entity Descriptions

We propose a new approach to build the entities' embeddings according to their descriptions. As in Fig. 1, we provide two examples of entity descriptions. They have been removed stop words, marked as [·] in the following:

1) Ithaca College: [is] [a] private college located [on] [the] South Hill [of] Ithaca,...;

2) New York: [is] [a] state [in] [the] Northeastern [and] Mid-Atlantic regions [of] [the] United States...

"Ithaca College" and *"New York"* are entities followed by their descriptions respectively. We formulate entity descriptions as $d_e:=\{w_1,w_2,\ldots,w_n\}$. d_e denotes the description of entity e. $\{w_1,w_2,\ldots,w_n\}$ is the set of words in entity description. $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_n$ are used to denote the embeddings of words $w_1,w_2,\ldots,w_n$ respectively, $w_1,w_2,\ldots,w_n\in\mathcal{W}$, $\mathcal{W}$ is used to denote the set of words. $\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_n\in\mathbf{W}$, $\mathbf{W}$ stands for the set of words' embeddings. n is the size of words set. In order to model entity description d_e, we use weighted bag of words (**WBOW**) model, and we use **TFIDF** to calculate their weights.

First, we calculate the term frequency ratio of word w_i in entity description d_e as follows:

$$\mathbf{TF}(w_i,d_e)=\frac{\text{term\_frequency}(w_i,d_e)}{\sum_{w\in d_e}\text{term\_frequency}(w,d_e)}\,,$$

where term_frequency(w,d_e) counts #times that word w occurs in entity descriptions d_e. Second, the inverse document (description) frequency of word w_i in the set of entity descriptions $\mathcal{D}$ is calculated as follows:

$$\mathbf{IDF}(w_i, \mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{d_e \in \mathcal{D} : w_i \in d_e\}|} ,$$

$\mathcal{D}$ is used to denote the set of descriptions. $|\mathcal{D}|$ is the total number of entity descriptions in the corpus $\mathcal{D}$, $|\{d_e \in \mathcal{D} : w_i \in d_e\}|$ is the number of entity descriptions where the word w_i appears (i.e., $\mathbf{TF}(w_i, d_e) \neq 0$). Then, the **TFIDF** of word w_i in entity description d_e in $\mathcal{D}$ is calculated as follows:

$$\mathbf{TFIDF}(w_i, d_e, \mathcal{D}) = \mathbf{TF}(w_i, d_e) \times \mathbf{IDF}(w_i, \mathcal{D}) .$$

We calculate the weight of word w_i in entity description d_e as follows:

$$\pi_i = \frac{\mathbf{TFIDF}(w_i, d_e, \mathcal{D})}{\sum_{w \in d_e} \mathbf{TFIDF}(w, d_e, \mathcal{D})} ,$$

with $\sum_{i \in 1,\ldots,n} \pi_i = 1$. Finally, we use **WBOW** model to calculate the zero-shot embedding of entity e according to its description d_e as follows:

$$\mathbf{e}_d = \sum_{i \in 1,\ldots,n} \pi_i \times \mathbf{w}_i , \tag{1}$$

with $\mathbf{w}_i \in \mathbb{R}^\kappa$, $\pi_i \in (0, 1)$.

For large scale KG and entity descriptions embedding, it requires that the approach for zero-shot embedding from description is less time-consuming and effective. The computational complexity of our method is $n \times \kappa$.

3.3 Jointly Embedding from KG and Entity Descriptions

We encode all the triples in KG to learn the embedding of In-KG entities, predicate relations and words. We consider that the model scoring function of the triple `(subject s, predicate p, object o)` depends on three factors: 1) the correlation between the subject entity and the predicate relation; 2) the correlation between the object entity and the predicate relation; 3) the correlation between the subject entity and the object entity. And all the three factors contribute to the final scoring function. A higher scoring value indicates a strong correlation. Thus, we have the model scoring function $\mathbf{G}(\mathtt{s}, \mathtt{p}, \mathtt{o})$ of the triple $(\mathtt{s}, \mathtt{p}, \mathtt{o})$ in KG as follows:

$$\mathbf{G}(\mathtt{s}, \mathtt{p}, \mathtt{o}) = \mathbf{f}_{sp}(\mathtt{s}, \mathtt{p}) + \mathbf{f}_{po}(\mathtt{p}, \mathtt{o}) + \mathbf{f}_{so}(\mathtt{s}, \mathtt{o}) , \tag{2}$$

where $\mathbf{f}_{sp}(\cdot)$, $\mathbf{f}_{po}(\cdot)$, and $\mathbf{f}_{so}(\cdot)$ denote the correlation between two arguments. For example, as the subject entity and the predicate relation, the correlation of them means their co-presence in knowledge base. We believe that the embeddings of them should be sort of similar if they often co-present in knowledge base. Of course, in general it is impossible that they are the same since each one (subject, or relation) would have correlations with other objects or other relations. Thus, we use inner product to measure their correlation, calculating $\mathbf{f}_{sp}(\cdot)$, $\mathbf{f}_{po}(\cdot)$ and $\mathbf{f}_{so}(\cdot)$. Note that it is not absolutely necessary to use inner product function to represent their interaction. The correlation takes high value means that they probably co-present in KG. Given three correlations take high value, the scoring function should be high value, so it would indicate that the triple should be true.

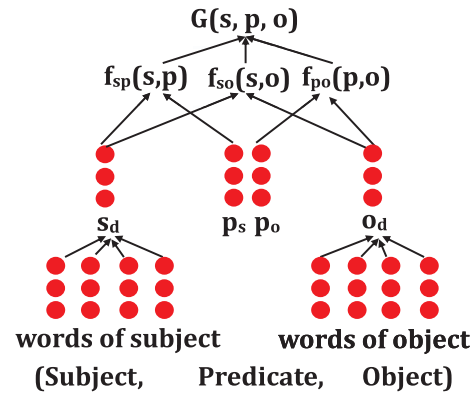

Fig. 2 The JointE model.

In addition, KG can be considered as a directed graph, i.e., the triple $(\mathtt{s}, \mathtt{p}, \mathtt{o})$ is different from its reverse triple $(\mathtt{o}, \mathtt{p}, \mathtt{s})$ in general. Thus, the scoring functions should be different. For distinguishing the different order information between them, we encode predicate relation as two embeddings $(\mathbf{p}_s, \mathbf{p}_o)$. The embeddings $\mathbf{p}_s$ and $\mathbf{p}_o$ interact with the embedding of subject entity **s** and object entity **o** respectively.

For subject entity `s` and object entity `o`, based on (1), we build their zero-shot embeddings according to their entity descriptions respectively as follows:

$$\mathbf{s}_d = \sum_{i \in 1,\ldots,n} \pi_{si} \times \mathbf{w}_{si} , \quad \mathbf{o}_d = \sum_{i \in 1,\ldots,m} \pi_{oi} \times \mathbf{w}_{oi} , \tag{3}$$

where n, m are the size of subject and object description's words set respectively. $\pi_{si}, \pi_{s2}, \ldots, \pi_{sn}$ and $\pi_{o1}, \pi_{o2}, \ldots, \pi_{om} \in (0, 1)$ are the weights of words for subject entity and object entity respectively, $\mathbf{w}_{si}, \mathbf{w}_{s2}, \ldots, \mathbf{w}_{sn}$ and $\mathbf{w}_{o1}, \mathbf{w}_{o2}, \ldots, \mathbf{w}_{om} \in \mathbb{R}^\kappa$ are the embeddings of words in descriptions of subject entity and object entity respectively. Combining (3) and (2), the scoring function of JointE model, as shown in Fig. 2, is provided as follows:

$$\begin{aligned} &\mathbf{G}(\mathtt{s}, \mathtt{p}, \mathtt{o}) := \mathbf{s}_d \cdot \mathbf{p}_s + \mathbf{p}_o \cdot \mathbf{o}_d + \mathbf{s}_d \cdot \mathbf{o}_d \\ &= (\sum_{i \in 1,\ldots,n} \pi_{si} \times \mathbf{w}_{si}) \cdot \mathbf{p}_s + \mathbf{p}_o \cdot (\sum_{i \in 1,\ldots,m} \pi_{oi} \times \mathbf{w}_{oi}) \\ &\quad + (\sum_{i \in 1,\ldots,n} \pi_{si} \times \mathbf{w}_{si}) \cdot (\sum_{i \in 1,\ldots,m} \pi_{oi} \times \mathbf{w}_{oi}) \end{aligned} \tag{4}$$

where $\mathbf{s}_d$, $\mathbf{o}_d$, $\mathbf{p}_s$, and $\mathbf{p}_o \in \mathbb{R}^\kappa$. κ is the dimension of the embeddings. The model returns a higher score if the triple $(\mathtt{s}, \mathtt{p}, \mathtt{o})$ is true in KG and a lower one otherwise.

4. Optimization

In this section, we propose a learning algorithm for training our model JointE in (4), in which the parameter set $\Theta = \{\mathbf{E}, \mathbf{P}_s, \mathbf{P}_o, \mathbf{W}\}$. $\mathbf{E}$ stands for the collection of all In-KG entities' embeddings. $\mathbf{P}_s$ and $\mathbf{P}_o$ denote the collections of predicate relations' embeddings. $\mathbf{W}$ stands for the set of

words' embeddings. After W is learnt in training phrase, and then those are used to build Out-of-KG entities' embeddings.

Objective function. We use contrastive max-margin (CMM) optimization criterion [3] to train our model (4). The main idea is that the model scoring function value of true knowledge triple in training set $\mathcal{T}$ should be larger than the corrupt one, the subject entity or object entity of which is replaced by a random one. Note that we do not replace both subject entity and object entity with random one at the same time. A triple will not be considered as a corrupt sample if it is already in training set $\mathcal{T}$. To learn embeddings $\Theta = \{\mathbf{E}, \mathbf{P}_s, \mathbf{P}_o, \mathbf{W}\}$, we minimize the hinge loss function $\mathbf{L}(\Theta)$ as follows:

$$\mathbf{L}(\Theta) = \sum_{(\mathsf{s},\mathsf{p},\mathsf{o})\in\mathcal{T}} \sum_{(\mathsf{s}',\mathsf{p},\mathsf{o}')\in\mathcal{T}'} \max\{0, \gamma - \mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o}) + \mathbf{G}(\mathsf{s}',\mathsf{p},\mathsf{o}')\} ,$$

where $\gamma > 0$ is a margin hyperparameter, $\mathbf{G}(\cdot)$ is the scoring function of JointE model, and

$$\mathcal{T}' := \{(\mathsf{s}',\mathsf{p},\mathsf{o}) | (\mathsf{s},\mathsf{p},\mathsf{o}) \in \mathcal{T} \cap \mathsf{s}' \in \mathcal{E} \cap \mathsf{s}' \neq \mathsf{s}\} \cup \{(\mathsf{s},\mathsf{p},\mathsf{o}') | (\mathsf{s},\mathsf{p},\mathsf{o}) \in \mathcal{T} \cap \mathsf{o}' \in \mathcal{E} \cap \mathsf{o}' \neq \mathsf{o}\} .$$

It is not absolutely necessary to use hinge loss function (e.g. sigmoid loss, etc). However, it is very common and normal to use hinge loss for learning embedding (like TransE, NTN, etc) such as our JointE model did.

Optimization. We use the stochastic gradient descent (SGD) algorithm for optimization. We initialize all the embeddings for In-KG entities, predicate relations and words $\{\mathbf{E}, \mathbf{P}_s, \mathbf{P}_o, \mathbf{W}\}$ with Gaussian distribution. Note that the set of words embedding $\mathbf{W}$ can be initialized by pre-trained word embedding result (e.g. Word2Vec learned on Wikipedia). In this paper, we do not use the pre-trained word embedding for initialization. We are going to learn the word embedding by JointE model from scratch with KG and entity descriptions. We perform the following procedure iteratively for a given number of iterations. First, we sample a small set (minibatch) of triples from the training set $\mathcal{T}$, and then for each positive triple in it, we construct a negative sample by replace the subject entity for object entity with random one. The parameters are then updated by taking a gradient descent step gradually. Algorithm 1 shows the detailed optimization algorithm. Note $[x]_+$ denotes the positive part of x (i.e. $[x]_+ := \max\{0, x\}$). $\mathcal{E}$ is used to denote sets of In-KG entities.

Algorithm 1 Learning JointE

Input: Training Set $\mathcal{T} = \{(\mathsf{s},\mathsf{p},\mathsf{o})\}$, margin hyperparameter γ, words' weight value set Π.
Output: The embeddings of all In-KG entities, predicates and words: $\Theta = \{\mathbf{E}, \mathbf{P}_s, \mathbf{P}_o, \mathbf{W}\}$.

Initialize
  $\mathbf{s}_d, \mathbf{o}_d \leftarrow \mathcal{N}(0,1)/10$ for each $\mathsf{s},\mathsf{o} \in \mathcal{E}, \mathbf{s}_d, \mathbf{o}_d \in \mathbb{R}^\kappa$
  $\mathbf{p}_s, \mathbf{p}_o \leftarrow \mathcal{N}(0,1)/10$ for each $\mathsf{p} \in \mathcal{R}, \mathbf{p}_s, \mathbf{p}_o \in \mathbb{R}^\kappa$
  $\mathbf{w} \leftarrow \mathcal{N}(0,1)/10$ for each $w \in \mathcal{W}, \mathbf{w} \in \mathbb{R}^\kappa$
Loop
  $\mathcal{T}_{batch} \leftarrow sample(\mathcal{T}, m)$//minibatch size m
  $\mathcal{H}_{batch} \in \phi$ //initialize training set as null
  for $(\mathsf{s},\mathsf{p},\mathsf{o}) \in \mathcal{T}_{batch}$ **do**
    $(\mathsf{s}',\mathsf{p},\mathsf{o}') \leftarrow sample\ \mathcal{T}$ //corrupted
    $\mathcal{H}_{batch} \leftarrow \mathcal{H}_{batch} \cup ((\mathsf{s},\mathsf{p},\mathsf{o}),(\mathsf{s}',\mathsf{p},\mathsf{o}'))$
  end for
  Update embeddings w.r.t.
  $\nabla\mathbf{L}(\Theta) = \sum_{\mathcal{H}_{batch}} \nabla[\gamma - \mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o}) + \mathbf{G}(\mathsf{s}',\mathsf{p},\mathsf{o}')]_+$, $\frac{\partial}{\partial \mathbf{s}_d}(\mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o})) \propto \mathbf{p}_\mathsf{s} + \mathbf{o}_\mathsf{d}$, $\frac{\partial}{\partial \mathbf{o}_\mathsf{d}}(\mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o})) \propto \mathbf{p}_\mathsf{o} + \mathbf{s}_\mathsf{d}$, $\frac{\partial}{\partial \mathbf{p}_s}(\mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o})) \propto \mathbf{s}_\mathsf{d}$, $\frac{\partial}{\partial \mathbf{p}_\mathsf{o}}(\mathbf{G}(\mathsf{s},\mathsf{p},\mathsf{o})) \propto \mathbf{o}_\mathsf{d}$, $\frac{\partial \mathbf{s}_d}{\partial \mathbf{w}_s} \propto \pi_s \cdot \mathbf{p}_s + \pi_s \cdot \pi_o \cdot \mathbf{w}_o$, $\frac{\partial \mathbf{o}_d}{\partial \mathbf{w}_o} \propto \mathbf{p}_o \cdot \pi_o + \pi_s \cdot \pi_o \cdot \mathbf{w}_s, \pi_s, \pi_o \in \Pi$
  $\mathbf{s}_d \leftarrow \mathbf{s}_d/\|\mathbf{s}_d\|$, $\mathbf{o}_d \leftarrow \mathbf{o}_d/\|\mathbf{o}_d\|$ for each entity embedding $\mathbf{s}_d, \mathbf{o}_d \in \mathbf{E}$
  $\mathbf{p}_s \leftarrow \mathbf{p}_s/\|\mathbf{p}_s\|$, $\mathbf{p}_o \leftarrow \mathbf{p}_o/\|\mathbf{p}_o\|$ for each predicate embedding $\mathbf{p}_s \in \mathbf{P}_s, \mathbf{p}_o \in \mathbf{P}_o$
End Loop

5. Experiments

The JointE model is evaluated on entity prediction for zero-shot embedding.

5.1 Datasets

Freebase is a large collaborative KG of general facts, currently including around 1.2 billion triples and more than 80 million entities. [3] extracted a subset from Freebase to build a dataset *FB15K* for knowledge base completion. [23] built a new dataset *FB20K* by taking *FB15K* as the seed and sharing the same predicate relations. All entities in Freebase which have predicate relations with entities in *FB15K* are selected as candidates. Then new entities from those candidates with rich descriptions are selected randomly. The average number of words in description is 69 after preprocessing, and the longest description contains 343 words. We use *FB20K* to simulate a zero-shot scenario that all entities in *FB15K* are considered as In-KG entities which can be learned through training, while 5,109 new-added entities are considered as Out-of-KG entities which are built from their descriptions. The training set in *FB20K* has 472,860 triples and 1,341 relations. *FB20K* has 3 types of test data: 1) $(d-e)$, the subject entity is a new entity (Out-of-KG) but the object entity is not new (In-KG); 2) $(e-d)$, the object entity is a new entity but the subject entity is not new; 3) $(d-d)$, both the subject entity and object entity are new entities.

WordNet is a large English lexical database, in which the entity corresponds to a concept (word sense) and the predicate relation defines the relation between two entities, such as the triple (_flint_NN_3, _part_of, _ wolverine_state_NN_1). The entities of WordNet are denoted by the concatenation of a word, its POS tag and a digital number. The number refers to its sense. E.g. "_flint_NN_3" encodes the third meaning of the noun "flint". [3] extracted a subset from WordNet, denoted by *WN*. We use *WN* as our data for experiments. We use Lucene (lucene.apache.org) to remove the stop words from entity descriptions in *WN*. To confirm that every entity has description for learning embedding, we remove the entities which have shorter than 3 words. Then,

Table 2 Statistics of the datasets used for evaluating JointE model by KGE in zero-shot scenario.

DATASETS		*FB20K*	*WN35K*
#In-KG Entities		14,904	28,307
#Out-of-KG Entities		5,019	6,891
#Predicates		1,341	18
#Words		68,547	28,601
#Train		472,860	71,088
#Test	$d-e$	18,753	16,798
	$e-d$	11,586	16,759
	$d-d$	151	4,000

we split the entities into two parts randomly: In-KG entities and Out-of-KG entities. We take all the triples in which the subject entity and object entity are In-KG entities from WN as training dataset. We extract all triples in which subject entity or object entity is a new entity (Out-of-KG) or both are new entities from WN as testing data. Thus, the testing dataset has 3 types: $(d-e)$, $(e-d)$ and $(d-d)$. The dataset is denoted by *WN35K*. The statistics of *FB20K* and *WN35K* are listed in Table 2.

5.2 Evaluation Metric

In the experiment, we use the ranking criteria [6] for evaluation. First, for each test triple, we remove the subject entity and replace it by each of the entities in turn. The model scoring function value (i.e. $\mathbf{G}(\mathtt{s}', \mathtt{p}, \mathtt{o})$) of the negative triples would be computed and then sorted by descending order in this paper. We can obtain the exact rank of the correct entity in the candidates. Similarly, we repeat the whole procedure while removing the object entity instead of the subject entity of the test triple. Finally, we use the proportion of correct entities ranked in the top 10 (Hits@10(%)) as the evaluation metric for comparison.

5.3 Baseline

We consider DKRL [23] as the compared model. The DKRL model is based on TransE. There are four types of the model: DKRL(CBOW), DKRL(CNN), DKRL(P-CBOW) and DKRL(P-CNN). In DKRL(CBOW) and DKRL(CNN), all entities use description-based representation. In DKRL(P-CBOW) and DKRL(P-CNN), entities in training set use structure-based representations. The author provides the experimental results on FB20K in original paper, so we directly use them as our baseline. In addition, we also run the DKRL code on another dataset WN35K for comparison. For comparison, we also use continuous bag-of-words (CBOW) (i.e., $\mathbf{e}_d = 1/n \sum_{i\in 1,\ldots,n} \mathbf{w}_i$) to encode entity descriptions, denoted as JointE(CBOW).

5.4 Parameter Setting

Like in original paper, we train the model DKRL with dimension κ in {50, 80, 100}, learning rate λ among {0.0005, 0.001,0.002}, and margin γ among {0.5, 1.0, 2.0}. The final configuration of DKRL(CBOW) is set as $\{\lambda = 0.001, \gamma = 1, \kappa = 50\}$ on *WN35K* by cross-validation on training data. For DKRL(CNN) encoder, we use 4-max-pooling and try different window size ℓ among {1, 2, 3} for different convolution layer. The dimension of word embedding n_w and feature map n_f are set among {50,80,100} and {50, 100, 150} respectively. The optimal configuration of DKRL(CNN) is $\{\lambda = 0.001, \gamma = 1, \kappa = 50, \ell = 2, n_w = 50, n_f = 50\}$ on *WN35K*. For determining appropriate hyperparameters for our model JointE, we select the learning rate λ_e(for entities' embedding), λ_p(for predicate relations' embedding) and λ_w (for words' embedding) among {0.001, 0.01, 0.1}, the margin γ among {1.0, 2.0, 5.0} and the embedding dimension κ in a range of {50, 100, 200} by cross-validation on test set. Finally, the configuration of both JointE(WBOW) and JointE(CBOW) are set up as $\{\kappa = 50, \lambda_e = 0.01, \lambda_p = 0.01, \lambda_w = 0.01, \gamma = 2\}$ on *FB15K*, and $\{\kappa = 50, \lambda_e = \lambda_p = \lambda_w = 0.01, \gamma = 2\}$ on *WN35K*. The #iteration of training is 1000.

Table 3 Evaluation results (Hits@10(%)) on entity prediction in zero-shot scenario on *FB20K*.

MODELS	$d-e$	$e-d$	$d-d$	Total
DKRL(P-CBOW)	26.5	20.9	67.2	24.6
DKRL(CBOW)	27.1	21.7	66.6	25.3
DKRL(P-CNN)	26.8	20.8	69.5	24.8
DKRL(CNN)	31.2	26.1	72.5	29.5
JointE(CBOW)	40.8	39.6	70	40.5
JointE(WBOW)	**44.4**	**39.9**	**85.6**	**42.9**

Table 4 Evaluation results (Hits@10(%)) on entity prediction in zero-shot scenario on *WN35K*.

MODELS	$d-e$	$e-d$	$d-d$	Total
DKRL(P-CBOW)	21.2	20.6	35.4	23.6
DKRL(CBOW)	22.5	21.7	34.5	25.2
DKRL(P-CNN)	23.5	21.2	37.6	23.1
DKRL(CNN)	24.9	24.3	39.2	28.6
JointE(CBOW)	33.5	34.0	46.5	34.3
JointE(WBOW)	**34.6**	**34.4**	**47.4**	**35.9**

5.5 Results of Entity Prediction

We evaluate the JointE model by predicting the subject entity and object entity of the triples in the testing data, while at least one entity in each testing triple is Out-of-KG entity. Table 3 and Table 4 show the evaluation results of our model against with the compared model DKRL on *FB20K* and *WN35K* respectively. From the results we observe that:

- The proposed model JointE(WBOW, CBOW) outperform the compared model DKRL(P-CBOW, CBOW, P-CNN, CNN) on *FB20K* and *WN35K*. More specifically, the JointE(WBOW) model improves 13.2%, 13.8% and 13.1% than DKRL(CNN) on $d-e$, $e-d$ and $d-d$ in *FB20K*, and also improves 9.7%, 10.1% and 8.2% in *WN35K* respectively. It improves 13.4% and 7.3% in total on *FB20K* and *WN35K* respectively. As same as TransE model [6], the DKRL model assumes translation relation between entities in embedding space.

Table 5 Examples of predicting object entity for KGE in zero-shot scenario. **Bold** indicates the true object entity.

	INPUT: SUBJECT ENTITIES AND PREDICATE RELATION	PREDICTION: OBJECT ENTITIES
$d-e$	Vincent Franklin /film/actor/film./film/performance/film	**The Illusionist**, Bright Star, The Bourne Identity, Saw IV
$e-d$	The University of Alabama /education/educational_institution/students_graduates ./education/education/student	**Rashi Bunny**, Ray Reach, Peter Riegert, Arthur Laurents
$d-d$	African ground squirrel /biology/organism_classification/lower_classifications	**Mountain ground squirrel**, The Yorkshire Terrier, Bulldog

Table 6 Examples of predicting subject entity for KGE in zero-shot scenario. **Bold** indicates the true subject entity.

	PREDICTION: SUBJECT ENTITIES	INPUT: PREDICATE RELATION AND OBJECT ENTITY
$d-e$	**Marty Adams**, Courteney Bass Cox, Carlos Auyero, Jeffrey Michael Tambor	/people/person/gender A male organism
$e-d$	Italy, **Malawi**, Mozambique, Saratoga County	/location/location/partially_contains Shire River
$d-d$	**Enlight Software**, Activision Blizzard, Nintendo Co., Ltd, Capcom	/cvg/cvg_publisher/games_published Glory of the Roman Empire

However, a drawback of TransE is that it can only model translating interactions of entities and relations in the triples, ignoring the intricate interactions among the items in the triple. For example, the first triple in Table 5, the translation model (TransE) of it is | Vincent Franklin + film - The Illusionist |, while the intricate interaction one (our model) is (Vincent Franklin, The Illusionist) + (Vincent Franklin, film) + (film, The Illusionist). And the translating interaction is a weak interaction between entities. Our proposed model JointE uses pairwise interaction to model triple. The pairwise interaction is a stronger interaction than translation interaction. We believe that the better performance of JointE than DKRL is due to the appropriate design of the model.

- JointE(WBOW) outperforms JointE(CBOW) on both *FB20K* and *WN35K*. More specifically, JointE(WBOW) improves 3.6%, 0.3% and 2.4% on $d-e$, $e-d$ and in total in *FB20K*, and improve 1.1%, 0.4%, 0.9% and 1.6% on $d-e$, $e-d$, $d-d$ and in total in *WN35K* respectively. It indicates the robustness of WBOW representations.
- Both JointE and DKRL perform better on $d-d$ than on $d-e$ and $e-d$, which indicates that the correlation between the Out-of-KG entities' embeddings is stronger than the correlation between the embeddings of the Out-of-KG entities and the In-KG entities. Note that the In-KG entities' embeddings are learnt directly from KG in training phase, while the Out-of-KG entities' embeddings (which are indirectly obtained) are built according to their descriptions after the training phase. We believe that it would result in a semantic gap between the Out-of-KG entities' embedding space and In-KG entities' embedding space, because the data which is used for learning In-KG entities embeddings and Out-of-KG entities embeddings are not the same, which can be demonstrated by this observation.

Case Study. Table 5 and Table 6 show the examples of predicting the *subject entity* and *object entity* respectively for KGE in the zero-shot scenario by JointE model, given the rest of the triple in testing data of *FB20K*. Given the input, the predicting results top-N (N = 4) are listed in order. The exact true answer is marked bold. We can see from the tables that most bold true subject or object are top ranked, which demonstrates the predicting capabilities of JointE model. However, given the predicate relation "partially contains" and object entity "Shire River", we can find the bold correct subject entity "Malawi" is not top-ranked, but all the predicted subject entity answer listed are countries' names. It indicates that even if the true fact is not always top-ranked, the predicted results can still reflect common-sense.

6. Conclusion

In this paper, we aim at extending knowledge graph in the zero-shot scenario, while most of the traditional approaches cannot deal with this issue, since they only learn the representation of In-KG entities and have no representation for unseen entities (or Out-of-KG entities). We propose a novel model (JointE) to jointly learn KG and entity descriptions embeddings to extend KG by adding relations with Out-of-KG entities. The JointE model builds representations for Out-of-KG entities using their descriptions. We evaluate the proposed model on two real datasets by entity prediction in the zero-shot scenario, and the experimental results show the effectiveness of the proposed model.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 61300080, No. 61273217, No. 61671078, the 111 Project under Grant No. B08004 and FP7 MobileCloud Project under Grant No. 612212. The au-

thors are partially supported by 360 Innovation, BUPT Excellent Ph.D. Students Foundation.

References

[1] G.A. Miller, "Wordnet: a lexical database for english," Communications of the ACM, vol.38, no.11, pp.39–41, 1995.
[2] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp.1247–1250, 2008.
[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," Proceedings of Advances in Neural Information Processing Systems, pp.2787–2795, 2013.
[4] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," Proceedings of the 25th Conference on Artificial Intelligence (AAAI), 2011.
[5] R. Socher, D. Chen, C.D. Manning, and A.Y. Ng, "Reasoning with neural tensor networks for knowledge base completion," Proceedings of Advances in Neural Information Processing Systems, pp.926–934, 2013.
[6] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "A semantic matching energy function for learning with multi-relational data," Machine Learning, vol.94, no.2, pp.233–259, 2014.
[7] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," Proceedings of the Conference on Artificial Intelligence (AAAI), pp.1112–1119, 2014.
[8] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," Proceedings of the Conference on Artificial Intelligence (AAAI), pp.2181–2187, 2015.
[9] Y. Zhao, S. Gao, P. Gallinari, and J. Guo, "Knowledge base completion by learning pairwise-interaction differentiated embeddings," Data Mining and Knowledge Discovery, vol.29, no.5, pp.1486–1504, 2015.
[10] R. Socher, M. Ganjoo, C.D. Manning, and A.Y. Ng, "Zero-shot learning through cross-modal transfer," Proceedings of Advances in Neural Information Processing Systems, pp.935–943, 2013.
[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning, Book in preparation for MIT Press, 2016.
[12] D. Zhang, B. Yuan, D. Wang, and R. Liu, "Joint semantic relevance learning with text data and graph knowledge," ACL-IJCNLP, pp.32–40, 2015.
[13] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," EMNLP, pp.1591–1601, 2014.
[14] Y. Bengio, H. Schwenk, J.S. Senecal, F. Morin, and J.L. Gauvain, "Neural probabilistic language models," Innovations in Machine Learning, pp.137–186, 2006.
[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," In Proceedings of the 25th international conference on Machine learning, pp.160–167, 2008.
[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," The Journal of Machine Learning Research, vol.12, pp.2493–2537, 2011.
[17] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," In Proceedings of the 24th international conference on Machine learning, pp.641–648, 2007.
[18] A. Mnih and G. Hinton, "A scalable hierarchical distributed language model," Proceedings of Advances in neural information processing systems, pp.1081–1088, 2009.
[19] T. Mikolov, M. Karafiat, L. Burget, and J. Cernock, "Recurrent neural network based language model," INTERSPEECH, vol.2, p.3, 2010.
[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv1301.3781, 2013.
[21] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Proceedings of Advances in neural information processing systems, pp.3111–3119, 2013.
[22] E.H. Huang, R. Socher, C.D. Manning, and A.Y. Ng, "Improving word representations via global context and multiple word prototypes," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp.873–882, 2012.
[23] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," Proceeddings of the Association for Advancement of Artificial Intelligence, 2016.
[24] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," Proceedings of Advances in Neural Information Processing Systems, pp.3464–3472, 2014.
[25] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1629–1636, 2014.
[26] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.819–826, 2013.
[27] X.N. Lam, T. Vu, T.D. Le, and A.D. Duong, "Addressing cold-start problem in recommendation systems," In Proceedings of the 2nd international conference on Ubiquitous information management and communication, pp.208–211, 2008.
[28] S.-T. Park and W. Chu, "Pairwise preference regression for cold-start recommendation," In Proceedings of the third ACM conference on Recommender systems, pp.21–28, 2009.
[29] K. Zhou, S.-H. Yang, and H. Zha, "Functional matrix factorizations for cold-start recommendation," In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp.315–324, 2011.

Yu Zhao received the B.S. degree from the Southwest Jiaotong University (SWJTU), Sichuan, in 2006 and the M.S. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, in 2011. He is currently pursuing the Ph.D. degree with the School of Information and Communication, BUPT. He has visited in LIP6, Universit Pierre et Marie Curie (UPMC), Paris, France, from May to August 2015. He is visiting in Department of Computer Science, University of Rochester (UR), Rochester, USA from Sep. 2015 to Mar. 2017. His research interests include natural language processing, machine learning, recommendation system, etc.

Sheng Gao has been an Assistant Professor with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, since 2012. He received the bachelor's and master's degrees from BUPT in 2003 and 2006, respectively, and the Ph.D. degree from Universite Pierre et Marie CURIE (Paris 6), Paris, France, in 2011. His current research interests include machine learning, data mining, information recommendation, and social network analysis. He has published over 20 academic papers on world-wide famous journals or conferences, including ISI, WWW, CIKM, and ECML.

Patrick Gallinari has been a Professor with Pierre et Marie Curie, Paris 6 University, Paris, France, since 1992, and the Director of the Computer Science Laboratory, LIP6, since 2005. He received the Ph.D. degree in computer science from the University of Compiegne, Compiegne, France, in 1985. His current research interests include machine learning applications and information retrieval.

Jun Guo is a Professor, Ph.D. Supervisor, Vice-President with the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, and the Dean of School of Information and Communication Engineering, BUPT, the Director of Pattern Recognition and Intelligent System Laboratory. He received the bachelor's and master's degrees from BUPT in 1982 and 1985, respectively, and the Ph.D. degree from Tohoku Gakuin University, Tohoku, Japan, in 1993. His current research interests include cross media information retrieval, web public sentiment information analysis, and network management and control. He is the person responsible for many projects funded by national 863 high-tech and national natural science foundation of China. He has published more than 100 papers on international journal and conference, including SCIENCE, Nature Online Magazine of Scientic Reports, and IEEE TPAMI.