Learning Embedding Representations for Knowledge Inference on Imperfect and Incomplete Repositories

Miao Fan†, Qiang Zhou and Thomas Fang Zheng‡
CSLT, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, China.
†fanmiao.cslt.thu@gmail.com,‡fzheng@tsinghua.edu.cn

***Abstract*—This paper considers the problem of knowledge inference on large-scale *imperfect* repositories with *incomplete* coverage by means of embedding entities and relations at the first attempt. We propose *IIKE* (Imperfect and Incomplete Knowledge Embedding), a probabilistic model which measures the probability of each belief, i.e. $\langle h, r, t\rangle$, in large-scale knowledge bases such as NELL and Freebase, and our objective is to learn a better low-dimensional vector representation for each entity (h and t) and relation (r) in the process of minimizing the loss of fitting the corresponding confidence given by machine learning (NELL) or crowdsouring (Freebase), so that we can use $||\mathbf{h}+\mathbf{r}-\mathbf{t}||$ to assess the plausibility of a belief when conducting inference. We use subsets of those inexact knowledge bases to train our model and test the performances of *link prediction* and *triplet classification* on ground truth beliefs, respectively. The results of extensive experiments show that *IIKE* achieves significant improvement compared with the baseline and state-of-the-art approaches.**

***Keywords*-Freebase; NELL; knowledge embedding; link prediction; triplet classification**

I. Introduction

The explosive growth in the number of web pages has drawn much attention to the study of information extraction [16] in recent decades. The aim of this is to distill unstructured online texts, so that we can store and exploit the distilled information as structured knowledge. Thanks to the long-term efforts made by experts, crowdsouring and even machine learning techniques, several web-scale knowledge repositories have been built, such as Wordnet, Freebase and NELL, and most of them contain tens of millions of extracted beliefs which are commonly represented by triplets, i.e. $\langle head\_entity, relation, tail\_entity\rangle$.

Although we have gathered colossal quantities of beliefs, state of the art work in the literature [20] reported that in this field, our knowledge bases are far from complete. For instance, nearly 97% persons in Freebase have unknown parents. To populate incomplete knowledge repositories, a large proportion of researchers follow the classical approach by extracting knowledge from texts [21], [1], [13]. For example, they explore ideal approaches that can automatically generate a precise belief like $\langle Madrid, capital\_city\_of, Spain\rangle$ from the sentence "*Madrid is the capital and largest city of Spain.*" on the web. However, even cutting-edge research [6] could not satisfy the demand of web-scale deployment, due to the diversification of natural language expression. Moreover, many implicit relations between two entities which are not recorded by web texts still need to be mined.

Therefore, some recent studies focus on inferring undiscovered beliefs based on the knowledge base itself without using extra web texts. One representative idea is to consider the whole repository as a graph where entities are nodes and relations are edges. The canonical approaches [15], [9], [10], [7] generally conduct *relation-specific* random walk inference based on the *local connectivity patterns* learnt from the imperfect knowledge graph. An alternative paradigm aims to perform *open-relation* inference via embedding all the elements, including entities and relations, into low-dimensional vector spaces. The proposed methods [18], [8], [2], [3], [17], [19] show promising performance, however, by means of learning from ground-truth training knowledge.

This paper thus contributes a probabilistic knowledge embedding model called *IIKE*[1] to measure the probability of each triplet, i.e. $\langle h, r, t\rangle$, and our objective is to learn a better low-dimensional vector representation for each entity (h and t) and relation (r) in the process of minimizing the loss of fitting the corresponding confidence given by machine learning (NELL) or crowdsouring (Freebase). To the best of our knowledge, *IIKE* is the first approach that attempts to learn *global connectivity patterns* for *open-relation* inference on imperfect and incomplete knowledge bases. In order to prove the effectiveness of the model, we conduct experiments on two tasks involved in knowledge inference, *link prediction* and *triplet classification*, using the two repositories mentioned above. Inexact beliefs are used to train our model, and we test the performance on ground truth beliefs. Results show that *IIKE* outperforms the other cutting-edge approaches on both different types of knowledge bases.

[1]It is short for Imperfect and Incomplete Knowledge Embedding.

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.17

II. Related Work

We group recent research work related to self-inferring new beliefs based on knowledge repositories without extra texts into two categories, graph-based inference models [15], [9], [10], [7] and embedding-based inference models [18], [8], [2], [3], [17], and describe the principal differences between them,

- *Symbolic representation v.s. Distributed representation*: Graph-based models regard the entities and relations as atomic elements, and represent them in a symbolic framework. In contrast, embedding-based models explore distributed representations via learning a low-dimensional continuous vector representation for each entity and relation.
- *Relation-specific v.s. Open-relation*: Graph-based models aim to induce rules or paths for a specific relation first, and then infer corresponding new beliefs. On the other hand, embedding-based models encode all relations into the same embedding space and conduct inference without any restriction on some specific relation.

A. Graph-based Inference

Graph-based inference models generally learn the representation for specific relations from the knowledge graph.

N-FOIL [15] learns first order Horn clause rules to infer new beliefs from the known ones. So far, it has helped to learn approximately 600 such rules. However, its ability to perform inference over large-scale knowledge repositories is currently still very limited.

PRA [9], [10], [7] is a data-driven random walk model which follows the paths from the head entity to the tail entity on the local graph structure to generate non-linear feature combinations representing the labeled relation, and uses logistic regression to select the significant features which contribute to classifying other entity pairs belonging to the given relation.

B. Embedding-based Inference

Embedding-based inference models usually design various scoring functions $f_r(h,t)$ to measure the plausibility of a triplet $\langle h,r,t\rangle$. The lower the dissimilarity of the scoring function $f_r(h,t)$ is, the higher the compatibility of the triplet will be.

Unstructured [3] is a naive model which exploits the occurrence information of the head and the tail entities without considering the relation between them. It defines a scoring function $||\mathbf{h}-\mathbf{t}||$, and this model obviously can not discriminate a pair of entities involving different relations. Therefore, *Unstructured* is commonly regarded as the baseline approach.

Distance Model (SE) [2] uses a pair of matrices (W_{rh}, W_{rt}), to characterize a relation r. The dissimilarity of a triplet is calculated by $||W_{rh}\mathbf{h}-W_{rt}\mathbf{t}||_1$. As pointed out by Socher et al. [17], the separating matrices W_{rh} and W_{rt} weaken the capability of capturing correlations between entities and corresponding relations, even though the model takes the relations into consideration.

Single Layer Model, proposed by Socher et al. [17] thus aims to alleviate the shortcomings of the *Distance Model* by means of the nonlinearity of a single layer neural network $g(W_{rh}\mathbf{h}+W_{rt}\mathbf{t}+\mathbf{b}_r)$, in which $g=tanh$. The linear output layer then gives the scoring function: $\mathbf{u}_r^T g(W_{rh}\mathbf{h}+W_{rt}\mathbf{t}+\mathbf{b}_r)$.

Bilinear Model [18], [8] is another model that tries to fix the issue of weak interaction between the head and tail entities caused by *Distance Model* with a relation-specific bilinear form: $f_r(h,t)=\mathbf{h}^T W_r\mathbf{t}$.

Neural Tensor Network (NTN) [17] designs a general scoring function: $f_r(h,t)=\mathbf{u}_r^T g(\mathbf{h}^T W_r\mathbf{t}+W_{rh}\mathbf{h}+W_{rt}\mathbf{t}+\mathbf{b}_r)$, which combines the *Single Layer Model* and the *Bilinear Model*. This model is more expressive as the second-order correlations are also considered into the nonlinear transformation function, but the computational complexity is rather high.

TransE [3] is a canonical model different from all the other prior arts, which embeds relations into the same vector space of entities by regarding the relation r as a translation from h to t, i.e. $\mathbf{h}+\mathbf{r}=\mathbf{t}$. It works well on the beliefs with ONE-TO-ONE mapping property but performs badly on multi-mapping beliefs. Given a series of facts associated with a ONE-TO-MANY relation r, e.g. $(h,r,t_1),(h,r,t_2),...,(h,r,t_m)$, *TransE* tends to represent the embeddings of entities on MANY-side extremely the same with each other and hardly to be discriminated.

TransH [19] is the state of the art approach as far as we know. It improves *TransE* by modeling a relation as a hyperplane, which makes it more flexible with regard to modeling beliefs with multi-mapping properties.

Even though the prior arts of knowledge embedding are promising when conducting *open-relation* inference on large-scale bases, the stage they stand on is made of ground-truth beliefs. The model *IIKE* that we have proposed belongs to the embedding-based community, but firstly tackles the problem with knowledge inference based on imperfect and incomplete repositories. Nevertheless, we compare our approach with the methods mentioned above, and assess the performance with both the dataset and the metrics they have used as part of the extensive experiments.

III. Model

The plausibility of a belief $\langle h,r,t\rangle$ can be regarded as the joint probability of the head entity h, the relation r and the tail entity t, namely $Pr(h,r,t)$. Similarly, $Pr(h|r,t)$ stands for the conditional probability of predicting h given r and t. We assume that $Pr(h,r,t)$ is collaboratively influenced by $Pr(h|r,t)$, $Pr(r|h,t)$ and $Pr(t|h,r)$, and more specifically it equals to the geometric

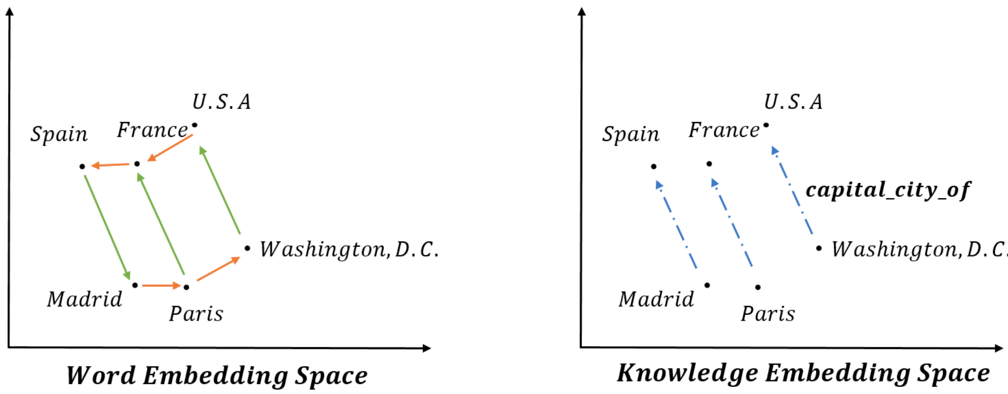

Figure 1. The result of vector calculation in the word embedding space: $\mathbf{v}_{Madrid}-\mathbf{v}_{Spain}+\mathbf{v}_{France}\approx\mathbf{v}_{Paris}$ and $\mathbf{v}_{Madrid}-\mathbf{v}_{Spain}+\mathbf{v}_{U.S.A}\approx\mathbf{v}_{Washington,D.C.}$ The most possible reason of $\mathbf{v}_{Spain}-\mathbf{v}_{Madrid}\approx\mathbf{v}_{France}-\mathbf{v}_{Paris}$ and $\mathbf{v}_{Spain}-\mathbf{v}_{Madrid}\approx\mathbf{v}_{U.S.A}-\mathbf{v}_{Washington,D.C.}$, is that $capital\_city\_of$ is the shared relation. In other words, $\mathbf{h}_{Madrid}+\mathbf{r}_{capital\_city\_of}\approx\mathbf{t}_{Spain}$, if the belief $\langle Madrid, capital\_city\_of, Spain\rangle$ is plausible.

mean of $Pr(h|r,t)Pr(r|h,t)Pr(t|h,r)$, which is shown in the subsequent equation,

$$Pr(h,r,t)=\sqrt[3]{Pr(h|r,t)Pr(r|h,t)Pr(t|h,r)}. \quad (1)$$

Given r and t, there are multiple choices of h' which may appear as the head entity. Therefore, if we use E_h to denote the set of all the possible head entities given r and t, $Pr(h|r,t)$ can be defined as

$$Pr(h|r,t)=\frac{\exp^{D(h,r,t)}}{\sum_{h'\in E_h}\exp^{D(h',r,t)}}, \quad (2)$$

The other factors, i.e. $Pr(r|h,t)$ and $Pr(r|h,t)$, are defined accordingly by slightly revising the normalization terms as shown in Equation (3) and (4), in which R and E_t represents the set of relations and tail entities, respectively.

$$Pr(r|h,t)=\frac{\exp^{D(h,r,t)}}{\sum_{r'\in R}\exp^{D(h,r',t)}}. \quad (3)$$

$$Pr(t|h,r)=\frac{\exp^{D(h,r,t)}}{\sum_{t'\in E_t}\exp^{D(h,r,t')}}. \quad (4)$$

The last function that we do not explain in Equation (2), (3) and (4) is $D(h,r,t)$. Inspired by somewhat surprising patterns learnt from word embeddings [12] illustrated by Figure 1, the result of word vector calculation, for instance $\mathbf{v}_{Madrid}-\mathbf{v}_{Spain}+\mathbf{v}_{France}$, is closer to $\mathbf{v}_{Paris}$ than to any other words [11]. If we study the example mentioned above, the most possible reason $\mathbf{v}_{Spain}-\mathbf{v}_{Madrid}\approx\mathbf{v}_{France}-\mathbf{v}_{Paris}$, is that $capital\_city\_of$ is the relation between $Madrid$ and $Spain$, and so is $Paris$ and $France$. In other words, $\mathbf{h}_{Madrid}+\mathbf{r}_{capital\_city\_of}\approx\mathbf{t}_{Spain}$, if the belief is plausible. Therefore, we define $D(h,r,t)$ as follows to calculate the dissimilarity between $\mathbf{h}+\mathbf{r}$ and $\mathbf{t}$ using L_1 or L_2 norm, and set b as the bias parameter.

$$D(h,r,t)=-||\mathbf{h}+\mathbf{r}-\mathbf{t}||+b. \quad (5)$$

So far, we have already modeled the probability of a belief, i.e. $Pr(h,r,t)$. On the other hand, some imperfect repositories, such as NELL, which is automatically built by machine learning techniques [5], assign a confidence score ($[0.5-1.0]$) to evaluate the plausibility of the corresponding belief. Therefore, we define the cost function $\mathcal{L}$ shown in Equation (6), and our objective is to learn a better low-dimensional vector representation for each entity and relation while continuously minimizing the total loss of fitting each belief $\langle h,r,t,c\rangle$ in the training set Δ to the corresponding confidence c.

$$\begin{aligned}\arg\min_{h,r,t}\ \ \mathcal{L}&=\sum_{\langle h,r,t,c\rangle\in\Delta}\frac{1}{2}(\log Pr(h,r,t)-\log c)^2\\&=\sum_{\langle h,r,t,c\rangle\in\Delta}\frac{1}{2}\{\frac{1}{3}[\log Pr(h|r,t)+\log Pr(r|h,t)\\&+\log Pr(t|h,r)]-\log c)\}^2.\end{aligned} \quad (6)$$

IV. Algorithm

To search for the optimal solution of Equation (6), we use *Stochastic Gradient Descent* (SGD) to update the embeddings of entities and relations in iterative fashion. However, it is cost intensive to directly compute the normalization terms in $Pr(h|r,t)$, $Pr(r|h,t)$ and $Pr(t|h,r)$. Enlightened by the work of Mikolov et al. [11], we have found an efficient approach that adopts negative sampling to transform the conditional probability functions, i.e. Equation (2), (3) and (4), to the binary classification problem, as shown in the subsequent equations,

$$\begin{aligned}&\log Pr(h|r,t)\approx\log Pr(1|h,r,t)\\&+\sum_{i=1}^{k}\mathbb{E}_{h'_i\sim Pr(h'\in E_h)}\log Pr(0|h'_i,r,t),\end{aligned} \quad (7)$$

$$
\begin{aligned}
&\log Pr(r|h,t) \approx \log Pr(1|h,r,t)\\
&+ \sum_{i=1}^{k} \mathbb{E}_{r'_i \sim Pr(r' \in R)} \log Pr(0|h, r'_i, t),
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
&\log Pr(t|h,r) \approx \log Pr(1|h,r,t)\\
&+ \sum_{i=1}^{k} \mathbb{E}_{t'_i \sim Pr(t' \in E_t)} \log Pr(0|h, r, t'_i),
\end{aligned} \tag{9}
$$

in (7), (8), and (9), we sample k negative beliefs and discriminate them from the positive case. For the simple binary classification problem mentioned above, we choose the logistic function with the offset ϵ shown in Equation (10) to estimate the probability that the given belief $\langle h, r, t\rangle$ is correct:

$$
Pr(1|h,r,t) = \frac{1}{1+\exp^{-D(h,r,t)}} + \epsilon. \tag{10}
$$

We also display the framework of the learning algorithm of *IIKE* in pseudocode as follows,

Algorithm 1 The Learning Algorithm of **IIKE**

Input:

Training set $\Delta = \{(h, r, t, c)\}$, entity set E, relation set R; dimension of embeddings d, number of negative samples k, learning rate α, convergence threshold η, maximum epoches n.

/*Initialization*/
foreach $\mathbf{r} \in R$ **do**
  $\mathbf{r} :=$ Uniform$(\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}})$
  $\mathbf{r} := \frac{\mathbf{r}}{|\mathbf{r}|}$
end foreach
foreach $\mathbf{e} \in E$ **do**
  $\mathbf{e} :=$ Uniform$(\frac{-6}{\sqrt{d}}, \frac{6}{\sqrt{d}})$
  $\mathbf{e} := \frac{\mathbf{e}}{|\mathbf{e}|}$
end foreach
/*Training*/
$i := 0$
while $Rel.loss > \eta$ and $i < n$ **do**
  foreach $\langle h, r, t\rangle \in \Delta$ **do**
    foreach $j \in$ range(k) **do**
      Negative sampling: $\langle h'_j, r, t\rangle \in \Delta'_h$
      /*Δ'_h is the set of k negative beliefs replacing h*/
      Negative sampling: $\langle h, r'_j, t\rangle \in \Delta'_r$
      /*Δ'_r is the set of k negative beliefs replacing r*/
      Negative sampling: $\langle h, r, t'_j\rangle \in \Delta'_t$
      /*Δ'_t is the set of k negative beliefs replacing t*/
    end foreach
    $\sum_{h,r,t,h',r',t'} \nabla \frac{1}{2}(\log Pr(h,r,t) - \log c)^2$
    /*Updating embeddings of $\langle h, r, t\rangle \in \Delta$, $\langle h', r, t\rangle \in \Delta'_h$, $\langle h, r', t\rangle \in \Delta'_r$, $\langle h, r, t'\rangle \in \Delta'_t$ with α and the batch gradients derived from Equation (7), (8), (9) and (10).*/
  end foreach
  i++
end while

Output:

All the embeddings of h, t and r, where $h, t \in E$ and $r \in R$.

V. Experiments

Embedding the entities and relations into low-dimensional vector spaces facilitates several classical knowledge inference tasks, such as *link prediction* and *triplet classification*. More specifically, link prediction performs inference via predicting a ranked list of missing entities or relations given the other two elements of a triplet. For example, it can predict a series of t given h and r, or a bunch of h given r and t. And triplet classification is to discriminate whether a triplet $\langle h, r, t\rangle$ is correct or wrong.

Several recent research works [3], [17], [19] reported that they used subsets of Freebase (**FB**) data to evaluate their models and showed the performance on the above two tasks, respectively. In order to conduct solid experiments, we compare our model (*IIKE*) with many related studies including the baseline and cutting-edge approaches mentioned in Section 2.2. Moreover, we use a larger imperfect and incomplete dataset (**NELL**) to perform comparisons involving the same tasks to show the superior inference capability of *IIKE*, and have released this dataset for others to use.

We are also glad to share all the datasets, the source codes and the learnt embeddings for entities and relations, which can be freely downloaded from http://pan.baidu.com/s/1mgxGbg8.

A. Link prediction

One of the benefits of knowledge embedding is that we can apply simple vector calculations to many reasoning tasks, and link prediction is a valuable task that contributes to completing the knowledge graph. With the help of knowledge embeddings, if we would like to tell whether the entity h has the relation r with the entity t, we just need to calculate the distance between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$. The closer they are, the more possibility the triplet $\langle h, r, t\rangle$ exists.

Table I
Statistics of the datasets used for link prediction task.

DATASET	FB15K	NELL
#(ENTITIES)	14,951	74,037
#(RELATIONS)	1,345	226
#(TRAINING EX.)	483,142	713,913
#(VALIDATING EX.)	50,000	7,296
#(TESTING EX.)	59,071	7,296

Table II

LINK PREDICTION RESULTS ON THE **FB15K** DATASET. WE COMPARED OUR PROPOSED *IIKE* WITH THE STATE-OF-THE-ART METHOD *TransH* AND OTHER PRIOR ARTS MENTIONED IN SECTION 2.2.

DATASET	FB15K			
METRIC	*MEAN RANK*		*MEAN HIT@10*	
	Raw	*Filter*	*Raw*	*Filter*
Unstructured [4]	1,074 / 14,951	979 / 14,951	4.5%	6.3%
RESCAL [14]	828 / 14,951	683 / 14,951	28.4%	44.1%
SE [2]	273 / 14,951	162 / 14,951	28.8%	39.8%
SME (LINEAR) [4]	274 / 14,951	154 / 14,951	30.7%	40.8%
SME (BILINEAR) [4]	284 / 14,951	158 / 14,951	31.3%	41.3%
LFM [8]	283 / 14,951	164 / 14,951	26.0%	33.1%
TransE [3]	243 / 14,951	125 / 14,951	34.9%	47.1%
TransH [19]	211 / 14,951	84 / 14,951	42.5%	58.5%
IIKE	**183** / 14,951	**70** / 14,951	**47.1%**	**59.7%**

Table III

LINK PREDICTION RESULTS ON THE **NELL** DATASET. WE COMPARED OUR PROPOSED *IIKE* WITH THE CUTTING-EDGE METHODS *TransH* AND *TransE*.

DATASET	NELL			
METRIC	*MEAN RANK*		*MEAN HIT@10*	
	Raw	*Filter*	*Raw*	*Filter*
TransE [3]	4,254 / 74,037	4,218 / 74,037	11.0%	12.3%
TransH [19]	3,469 / 74,037	**2,218** / 74,037	25.2%	**41.6%**
IIKE	**2,464** / 74,037	2,428 / 74,037	**37.3%**	38.2%

1) Datasets: Bordes et al. [3] released a large dataset (**FB15K**)[2], extracted from Freebase and constructed by crowdsourcing, in which each belief is a triplet without a confidence score. Therefore, we assign 1.0 to each training triplet by default. We have also identified a larger repository on the web named **NELL**[3] which is automatically built by machine learning techniques, and each triplet is labeled with a probability estimated by synthetic algorithms [5]. We reserve the beliefs with probability ranging (0.5 - 1.0], use the ground-truth (1.0) beliefs as the validating and testing examples, and train the models with the remains.

Table 1 shows the statistics of these two datasets. The scale of **NELL** dataset is larger than **FB15K** with many more entities but fewer relations, which may lead to the differences of tuning parameters[4].

2) Evaluation Protocol: For each testing triplet, all the other entities that appear in the training set take turns to replace the head entity. Then we get a bunch of candidate triplets associated with the testing triplet. The dissimilarity of each candidate triplet is firstly computed by various scoring functions, such as $||\mathbf{h}+\mathbf{r}-\mathbf{t}||$, and then sorted in ascending order. Finally, we locate the ground-truth triplet and record its rank. This whole procedure runs in the same way when replacing the tail entity, so that we can gain the mean results. We use 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. However, the results measured by those metrics are relatively inaccurate, as the procedure above tends 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.

3) Experimental Results: We compared *IIKE* with the state-of-the-art *TransH*, *TransE* and other models mentioned in Section 2.2 evaluated on **FB15K** and **NELL** . We tuned the parameters of each previous model[5] based on the validation set, and select the combination of parameters which leads to the best performance. The results of prior arts on **FB15K** are the same as those reported by Wang et al. [19]. For *IIKE*, we tried several combinations of parameters: $d = \{20, 50, 100\}$, $\alpha = \{0.1, 0.05, 0.01, 0.005, 0.002\}$, $b = \{7.0, 10.0, 15.0\}$ and $norm = \{L_1, L_2\}$, and finally chose $d = 50$, $\alpha = 0.002$, $b = 7.0$, $norm = L_2$ for the **FB15K** dataset, and $d = 100$, $\alpha = 0.001$, $b = 7.0$, $norm = L_1$ for the **NELL** dataset. Moreover, to make responsible comparisons between *IIKE* and the state-of-the-art approaches, we requested the authors of *TransH* to re-run their system on the **NELL** dataset and reported the best results. Table 2 demonstrates that *IIKE* outperforms all the prior arts, including the baseline model *Unstructured* [4], *RESCAL* [14], *SE* [2], *SME (LINEAR)* [4], *SME (BILINEAR)* [4], *LFM* [8] and *TransE* [3], and achieves significant improvements on the **FB15K** dataset, compared with the

[2]Related studies on this dataset can be looked up from the website https://www.hds.utc.fr/everest/doku.php?id=en:transe

[3]The whole dataset of NELL can be downloaded from http://rtw.ml.cmu.edu/rtw/resources

[4]It turns out that embedding models prefer a larger dimension of vector representations for the dataset with more entities, and L_1 *norm* for fewer relations.

[5]All the codes for the related models can be downloaded from https://github.com/glorotxa/SME

state-of-the-art *TransH* [19]. For the **NELL** dataset, *IIKE* performs stably on the evaluation metrics compared with *TransH* and *TransE*, as Table 3 shows that it improves by 28.9% in terms of *Raw Mean Rank*, and achieves comparable performance of *Filter Mean Rank* compared with *TransH*.

B. Triplet classification

Triplet classification is another inference related task proposed by Socher et al. [17] which focuses on searching a relation-specific threshold σ_r to identify whether a triplet $\langle h, r, t\rangle$ is plausible.

Table IV
STATISTICS OF THE DATASETS USED FOR TRIPLET CLASSIFICATION TASK.

DATASET	FB15K	NELL
#(ENTITIES)	14,951	74,037
#(RELATIONS)	1,345	226
#(TRAINING EX.)	483,142	713,913
#(VALIDATING EX.)	100,000	14,592
#(TESTING EX.)	118,142	14,582

1) Datasets: Wang et al. [19] constructed a standard dataset **FB15K** sampled from Freebase. Moreover, we build another imperfect and incomplete dataset, i.e. **NELL**, following the same principle that the head or the tail entity can be randomly replaced with another one to produce a negative triplet, but in order to build much tough validation and testing datasets, the principle emphasizes that the picked entity should once appear at the same position. For example, *(Pablo Picaso, nationality, American)* is a potential negative example rather than the obvious irrational *(Pablo Picaso, nationality, Van Gogh)*, given a positive triplet *(Pablo Picaso, nationality, Spanish)*, as *American* and *Spanish* are more common as the tails of *nationality*. And the beliefs in the training sets are the same as those used in triplet classification. Table 4 shows the statistics of the standard datasets that we used for evaluating models on the triplet classification task.

2) Evaluation Protocol: The decision strategy for binary classification is simple: if the dissimilarity of a testing triplet (h, r, t) computed by $f_r(h, t)$ is below the relation-specific threshold σ_r, it is predicted as positive, otherwise negative. The relation-specific threshold σ_r can be searched via maximizing the classification accuracy on the validation triplets which belong to the relation r.

3) Experimental Results: We use the best combination of parameter settings in the link prediction task: $d = 50$, $\alpha = 0.002$, $b = 7.0$, $norm = L_2$ for the **FB15K** dataset, and $d = 100$, $\alpha = 0.001$, $b = 7.0$, $norm = L_1$ for the **NELL** dataset, to generate the entity and relation embeddings, and learn the best classification threshold σ_r for each relation r. Compared with several of the latest approaches, i.e.

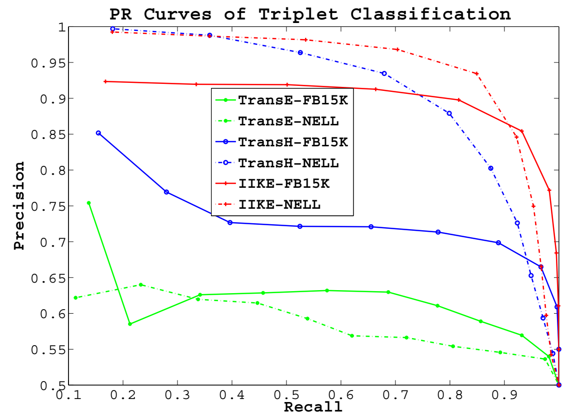

Figure 2. The comparison of precison-recall curves for triplet classification among the proposed *IIKE* (red lines), the state-of-the-art approaches *TransH* (blue lines) and *TransE* (green lines).

TransH [19], *TransE* [3] and *Neural Tensor Network (NTN)*[6] [17], the proposed *IIKE* approach still outperforms them, as shown in Table 5. We also drew the precision-recall curves which indicate the capability of global discrimination by ranking the distance of all the testing triplets, and Figure 2 shows that the AUC (Areas Under the Curve) of *IIKE* is much bigger than the other approaches.

Table V
THE ACCURACY OF TRIPLET CLASSIFICATION COMPARED WITH SEVERAL LATEST APPROACHES: *TransH*, *TransE* AND *NTN*.

DATASET	FB15K	NELL
NTN [17]	66.7%	-
TransE [3]	79.7%	82.4%
TransH [19]	80.2%	89.1%
IIKE	**91.1%**	**91.4%**

VI. CONCLUSION

We challenge the problem of knowledge inference on imperfect and incomplete repositories in this paper, and have produced an elegant probabilistic embedding model to tackle this issue at the first attempt by measuring the probability of a given belief $\langle h, r, t\rangle$. To efficiently learn the embeddings for each entity and relation, we also adopt the negative sampling technique to transform the original model and display the algorithm based on SGD to search the optimal solution. Extensive experiments on knowledge inference including *link prediction* and *triplet classification* show that our approach achieves significant improvement on two large-scale knowledge bases, compared with state-of-the-art and baseline methods.

[6] Socher et al. reported higher classification accuracy in [17] with word embeddings. In order to conduct a fair comparison, the accuracy of *NTN* reported in Table 5 is same with the EV (entity vectors) results in Figure 4 of [17].

We are pleased to see further improvements of the proposed model, which leaves open promising directions for the future work, such as taking advantage of the knowledge embeddings to enhance the studies of text summarization and open-domain question answering.

VII. Acknowledgments

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.

References

[1] Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Literature review for Language and Statistics II*, 2007.

[2] Antoine Bordes, Jason Weston, Ronan Collobert, Yoshua Bengio, et al. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795, 2013.

[4] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, 2014.

[5] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[6] Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y. Chang. Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[7] Matt Gardner, Partha Pratim Talukdar, Bryan Kisiel, and Tom M. Mitchell. Improving learning and inference in a large knowledge-base using latent syntactic cues. In *EMNLP*, pages 833–838. ACL, 2013.

[8] Rodolphe Jenatton, Nicolas Le Roux, Antoine Bordes, Guillaume Obozinski, et al. A latent factor model for highly multi-relational data. In *NIPS*, pages 3176–3184, 2012.

[9] Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[10] Ni Lao, Tom Mitchell, and William W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

[12] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. The Association for Computational Linguistics, 2013.

[13] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[14] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816, 2011.

[15] J. Ross Quinlan and R. Mike Cameron-Jones. Foil: A midterm report. In *Proceedings of the European Conference on Machine Learning*, ECML '93, pages 3–20, London, UK, UK, 1993. Springer-Verlag.

[16] Sunita Sarawagi. Information extraction. *Foundations and trends in databases*, 1(3):261–377, 2008.

[17] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934, 2013.

[18] Ilya Sutskever, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Modelling relational data using bayesian clustered tensor factorization. In *NIPS*, pages 1821–1828, 2009.

[19] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 1112–1119, 2014.

[20] Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *WWW*, 2014.

[21] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.