Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Joint entity–relation knowledge embedding via cost-sensitive learning*

Sheng-kang YU[1], Xue-yi ZHAO[1], Xi LI[2], Zhong-fei ZHANG[‡1]

([1]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

E-mail: shengkangyu@zju.edu.cn; xueyizhao@zju.edu.cn; xilizju@zju.edu.cn; zhongfei@zju.edu.cn

Received May 13, 2016; Revision accepted June 26, 2016; Crosschecked Nov. 22, 2017

Abstract: As a joint-optimization problem which simultaneously fulfills two different but correlated embedding tasks (i.e., entity embedding and relation embedding), knowledge embedding problem is solved in a joint embedding scheme. In this embedding scheme, we design a joint compatibility scoring function to quantitatively evaluate the relational facts with respect to entities and relations, and further incorporate the scoring function into the max-margin structure learning process that explicitly learns the embedding vectors of entities and relations using the context information of the knowledge base. By optimizing the joint problem, our design is capable of effectively capturing the intrinsic topological structures in the learned embedding spaces. Experimental results demonstrate the effectiveness of our embedding scheme in characterizing the semantic correlations among different relation units, and in relation prediction for knowledge inference.

Key words: Knowledge embedding; Joint embedding; Cost-sensitive learning
https://doi.org/10.1631/FITEE.1601255 **CLC number:** TP393

1 Introduction

Knowledge representation and inference in knowledge base are crucial and challenging problems in the age of big data. Generally, the pivots of the knowledge base are constituted by unary atomic units (i.e., entities), while the semantic structure of the knowledge base is constituted by higher-level and interactive atomic units (i.e., relations).

As a popular and powerful tool of knowledge representation and inference, knowledge embedding (Getoor and Mihalkova, 2011; Hoffmann *et al.*, 2011; Chang *et al.*, 2013) constructs a statistical model on a low-dimensional vector space to quantitatively evaluate these atomic units.

To learn effective embeddings, semantically similar units should be embedded near each other in the embedding space, and the main factor to consider is the context information of the knowledge base, as shown in Fig. 1. The knowledge base is embedded into the entity and relation embedding spaces. The embeddings of entities are restricted by relations among them while those of relations are restricted by a higher-order context. In each space, semantically close objects should be near each other. The context information exists at both entity and relation levels. At the entity level, the context refers to the relations connecting the entities. It determines the relative structure of entities and is important for knowledge representation. at the relation level, the context refers to the correlation among the relations. It has a higher order than that at the entity level and is helpful for inference. The knowledge embedding should be the model

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2015CB352302) and the National Natural Science Foundation of China (Nos. U1509206 and 61472353)

ORCID: Xi LI, http://orcid.org/0000-0003-3023-1662

of taking both the entity and relation levels into consideration, with the capability of representing the knowledge base by the coordinates and maintaining the major context information.

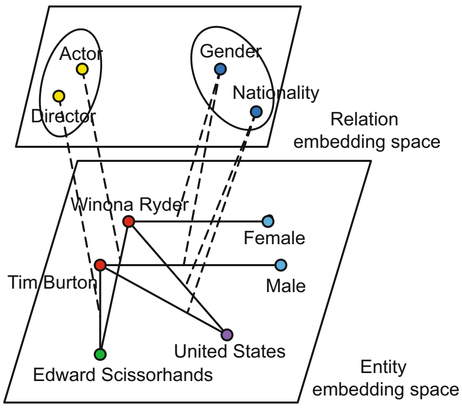

Fig. 1 An diagrammatic sketch of the joint knowledge embedding

Most conventional embedding work (Bordes *et al.*, 2011; 2013; 2014; Nickel *et al.*, 2011; Jenatton *et al.*, 2012; García-Durán *et al.*, 2015; Lin *et al.*, 2015) focuses on embedding at the entity level while the relations are not explicitly embedded, making it difficult to take advantages of the higher-order context information. In fact, the embedding at the relation level is also important, since relations contain a rich semantic content and the relationships among the relations reflect the intrinsic topology structure of the knowledge base. As a result, it is of great help to seek an embedding space and model the relations as embedding vectors in the space to measure their structure information, namely relation embedding.

In general, entity and relation embeddings are two important aspects of knowledge embedding. The relational facts in the knowledge base are determined by both entities and relations. Only by embedding the entities and relations well simultaneously can the knowledge base be well represented, since entities and relations are highly related by semantics. In the embedding process, entities are restricted by the relations, while relations are supported by entities. Hence, entity and relation embeddings are mutually correlated and complementary. Thus we consider knowledge embedding as a joint optimization problem, and propose our embedding model by connecting the entity and relation prediction tasks.

In our approach, we introduce a cost-sensitive learning scheme for joint knowledge embedding. We define an entity–relation–entity score function to evaluate the relational facts and optimize the entity and relation embedding vectors in a cost-sensitive learning framework. Besides, to capture the context information of the relations, a relationwise weighted score function is formulated in a max-margin way. The score function which reflects the topology of the entities, and the cost-sensitive learning framework which reflects the topology of the relations, together provide the jointly optimized embedding vectors of both entities and relations. As a result, our model involves the context information of the knowledge base and generally leads to semantically meaningful embedding results.

In summary, the main contributions of this work are summarized as follows:

1. We propose a joint embedding scheme for knowledge embedding, in which entities and relations are embedded to different spaces simultaneously, and the explicit embedding vectors reflect the topological structure of the knowledge base.

2. We design a cost-sensitive learning framework in the joint embedding scheme to involve the context information of the knowledge base in the optimization task and to solve the incompleteness problem of the knowledge base.

2 Multi-task knowledge embedding

2.1 Joint embedding of entities and relations

The structured content of knowledge bases is typically represented in a triplet form (Singhal, 2012; Waters, 2012). Typically, each triplet is a relational fact. It is composed of a head entity $\boldsymbol{e}_1$, a relation $\boldsymbol{r}$, and a tail entity $\boldsymbol{e}_2$. Therefore, a triplet is typically denoted as $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)$. Our joint embedding scheme aims to learn a low-dimensional embedding vector $\boldsymbol{e}$ for each entity and an embedding vector $\boldsymbol{r}$ for each relation according to the knowledge base. The embedding results should obey the semantic rules (i.e., the known triplets), and can predict unknown links which predict the correct relations between two given entities $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, by solving a relation embedding problem.

To learn effective embeddings, it is necessary to define an evaluation metric of the embedding vectors as the purpose of the optimization. In our joint embedding scheme, we define a score function $S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)$ to evaluate the plausibility of triplet $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)$ by exploiting the corresponding embedding vectors $\boldsymbol{e}_1$, $\boldsymbol{r}$, and $\boldsymbol{e}_2$. To connect the entity and the relation embedding spaces, we restrict the two spaces with the same dimension and define the score function as a vector product of the three embedding vectors in a triplet: the inner product of the entity embedding vectors and the relation embedding vectors for weighting. Thus the score function is written as

$$S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) = \boldsymbol{e}_1^{\mathrm{T}} \mathrm{diag}(\boldsymbol{r}) \boldsymbol{e}_2, \tag{1}$$

where $\mathrm{diag}(\boldsymbol{r})$ represents the diagonal matrix whose diagonal elements are vector $\boldsymbol{r}$. All the vectors are column vectors.

To satisfy the existing triplet facts, the correct relation between two given entities should always score higher than the incorrect relations, e.g, $S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) > S(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2)$, where $(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2)$ is a triplet which is not included in the knowledge base. Thus our model aims to satisfy the following inequalities:

$$\boldsymbol{e}_1^{\mathrm{T}} \mathrm{diag}(\boldsymbol{r}) \boldsymbol{e}_2 > \boldsymbol{e}_1^{\mathrm{T}} \mathrm{diag}(\boldsymbol{r}') \boldsymbol{e}_2, \tag{2}$$

where $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) \in T$, $(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \notin T$, and T represents the training set of the triplets.

To enhance the embedding result, a margin γ satisfying $S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) - S(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \geqslant \gamma$ is used for encouraging the discrimination between correct and incorrect triplets. Thus a ranking model is built with the following margin-based (Elisseeff and Weston, 2001) training objective:

$$\begin{aligned} &\min_{\boldsymbol{e},\boldsymbol{r},\boldsymbol{e}\in\mathcal{E},\boldsymbol{r}\in\mathcal{R}} \sum_{(\boldsymbol{e}_1,\boldsymbol{r},\boldsymbol{e}_2)\in T} \sum_{(\boldsymbol{e}_1,\boldsymbol{r}',\boldsymbol{e}_2)\notin T} \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'} + C \sum_{\boldsymbol{r}\in\mathcal{R}} ||\boldsymbol{r}||_2^2 \\ &\text{s.t.} \quad \begin{cases} S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) - S(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \geqslant \gamma - \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'}, \\ \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'} \geqslant 0, \end{cases} \end{aligned} \tag{3}$$

where $\mathcal{E}$ and $\mathcal{R}$ are the collections of entities and relations, respectively, and C is a hyperparameter of the regularization terms of the relation embeddings. In theory, the training objective should also include regularization terms of the relation embeddings. To simplify computation, we normalize the entity embedding vectors into unit vectors. Thus there are no regularization terms of entity embeddings. Since the norm of the relation embeddings is different from that of the entity embeddings, it is impossible to normalize the relation and the entity embeddings in the training objective simultaneously, and thus the regularization terms of the relation embeddings are maintained.

In the state-of-the-art entity embedding framework, negative samples are obtained mainly by randomly exchanging the head or tail entity of a triplet in the training set. To emphasize the importance of the relations, we generate negative samples by exchanging relations from a triplet. The experimental results demonstrate that it is helpful in improving the performance of the relation prediction.

2.2 Cost-sensitive learning

In Section 2.1, the optimization objective involves only the context information at the entity level, since the relationships among the relations are not considered. In the following, we further investigate the context information at the relation level in a cost-sensitive manner.

The most important context information at the relation level is the concurrence of the relations. Some relations are likely to occur at the same time between two entities, and some other relations cannot co-exist. In fact, problem (3) assigns equal weights to all pairs of relations, with the assumption that all the relations co-occur with the same probability. To take advantages of higher-order context information among the relations, we introduce a weight matrix $\boldsymbol{W}$ to measure the concurrence probability of each pair of relations. Each element of $\boldsymbol{W}$ reflects the plausibility of a relation to connect two entities if they are connected by another relation. When computing the losses of $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) \in T$ and $(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \notin T$, we add a weight $\boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}$ to the corresponding loss. If two entities are connected with relation $\boldsymbol{r}$ in reality, $\boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}$ should be low; otherwise, $\boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}$ should be high. Therefore, the objective of the cost-sensitive ranking system is formulated as

$$\begin{aligned} &\min_{\boldsymbol{e},\boldsymbol{r},\boldsymbol{W},\boldsymbol{e}\in\mathcal{E},\boldsymbol{r}\in\mathcal{R}} \sum_{(\boldsymbol{e}_1,\boldsymbol{r},\boldsymbol{e}_2)\in T} \sum_{(\boldsymbol{e}_1,\boldsymbol{r}',\boldsymbol{e}_2)\notin T} \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'} \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'} \\ &+ C \sum_{\boldsymbol{r}\in\mathcal{R}} ||\boldsymbol{r}||_2^2 \end{aligned} \tag{4}$$

$$\text{s.t.} \quad \begin{cases} S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) - S(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \geqslant \gamma - \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'}, \\ \xi_{\boldsymbol{t}\boldsymbol{r}\boldsymbol{r}'} \geqslant 0, \\ \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'} \geqslant 0, \\ \sum \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'} = \delta, \\ \delta > 0. \end{cases}$$

The cost-sensitive model involves high-order context information and enriches potentially correct triplets that are not included in the knowledge base. Compared with problem (3), the cost-sensitive model recognizes relations, which potentially co-occur with the correct relationships, and reduces the risk of taking a potentially correct triplet as a totally negative sample.

2.3 Model solving

Since problem (4) includes a large number of constraints, it is quite time-consuming to solve in the dual space to acquire the global optimized solutions. Thus we transform the objective to an unconstrained Lagrangian form in the original space, and solve it via the stochastic gradient descent method (Duchi *et al.*, 2011) with the mini-batch strategy. The training procedure is shown in Algorithm 1.

Algorithm 1 Training procedure of the joint embedding framework

Input: training set T, embedding dimension k, hyperparameter C, and constraint γ.

Output: embedding vector $\boldsymbol{e}_i$ for each entity and embedding vector $\boldsymbol{r}$ for each relation.

Initialize: initialize each $\boldsymbol{e}_i$ as a k-dimensional unit vector; initialize $\boldsymbol{r}$ as a non-negative k-dimensional vector; initialize $\boldsymbol{W}$ as a $k \times k$-dimensional non-negative matrix.

repeat
  Randomly select a mini-batch T' from the training set;
  Set all gradients to 0;
  for each $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) \in T'$ **do**
    for each $(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) \notin T$ **do**
      if $S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2) - S(\boldsymbol{e}_1, \boldsymbol{r}', \boldsymbol{e}_2) < \gamma$ **then**
        $\frac{\partial L}{\partial \boldsymbol{e}_1} = \frac{\partial L}{\partial \boldsymbol{e}_1} - \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}(\boldsymbol{r} - \boldsymbol{r}').*\boldsymbol{e}_2$;
        $\frac{\partial L}{\partial \boldsymbol{e}_2} = \frac{\partial L}{\partial \boldsymbol{e}_2} - \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}(\boldsymbol{r} - \boldsymbol{r}').*\boldsymbol{e}_1$;
        $\frac{\partial L}{\partial \boldsymbol{r}} = \frac{\partial L}{\partial \boldsymbol{r}} - \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}\boldsymbol{e}_1.*\boldsymbol{e}_2$;
        $\frac{\partial L}{\partial \boldsymbol{r}'} = \frac{\partial L}{\partial \boldsymbol{r}'} + \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}\boldsymbol{e}_1.*\boldsymbol{e}_2$;
        $\frac{\partial L}{\partial \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}} = \frac{\partial L}{\partial \boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}} + \gamma - \boldsymbol{e}_1^{\mathrm{T}}\mathrm{diag}(\boldsymbol{r})\boldsymbol{e}_2 + \boldsymbol{e}_1^{\mathrm{T}}\mathrm{diag}(\boldsymbol{r})\boldsymbol{e}_2$;
      end if
    end for
  end for
  Update $\boldsymbol{e}_i$, $\boldsymbol{r}_i$, and $\boldsymbol{W}_{\boldsymbol{r}\boldsymbol{r}'}$ by the gradient descent;
until the convergence condition is met.

Table 1 Datasets used in the experiments of relation prediction and triplet prediction

Dataset	Number of triplet sets				
	Relations	Entities	Train	Validation	Test
WordNet	11	13 503	34 162	7912	7913
FB13	13	24 999	65 864	15 772	15 772
FB15k	1389	14 951	350 159	100 000	142 213

3 Experiments

3.1 Datasets

We used two different knowledge base datasets, namely WordNet (Miller, 1995) and FreeBase (Bollacker *et al.*, 2008), to evaluate the performances of prediction. WordNet is a lexical database for English language. It groups English words into sets of synonyms and records a number of relations among these synonym sets or their members. To construct the triplets, we used the entities corresponding to word meanings, and lexical relations defined by the relationships among them. FreeBase is a large collaborative knowledge base of structured data harvested from many sources. In these datasets, entities include various objects such as real people, places, and organizations. Relations of the datasets are listed in an index form to express their hierarchical structure, such as '/music/group_member/membership' to distinguish the membership of bands from the membership of other organizations. Two subsets of FreeBase, FB13 and FB15k, were used in the experiments. Details of the sets are documented in Table 1.

3.2 Experimental tasks and evaluation protocol

The main experiments to evaluate the performance of our knowledge embedding scheme are the relation embedding distribution and relation prediction. We learned the entity and relation embedding vectors from the training sets, investigated the relation embedding distribution, and conducted relation prediction task on the test sets based on the learned embedding vectors.

The relation embedding distribution experiment is to indicate how well the relations is represented by the relation embedding vectors. To validate that our embedding model embeds semantically close relations as neighbors in the embedding space, we propose a visualization result by reducing the embedding results into a low-dimensional space, and show several examples of the nearest neighbors of relations. The dataset that we chose to conduct the relation embedding distribution experiment is the FB15k dataset, since it includes multiple relations. The compared approach is TransE (Bordes *et al.*, 2013), which embeds all the relations into vectors in one space.

The relation prediction experiment, which predicts the potential relations between two given entities, was conducted to indicate the inference ability of the embedding. Given $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, relation prediction aims to acquire a ranking list R which satisfies that the more forward $\boldsymbol{r}$ in R corresponds to the more potential $(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)$. The performances were evaluated by two generic metrics: MeanRank and Hits@3 (or Hits@10 on the FB15k dataset) following the conventional work (Bordes *et al.*, 2013; Wang *et al.*, 2014). MeanRank is the mean rank of correct relations. In the experiments, if there exist any correct relations ranked ahead the test correct relations, we filtered them to ensure that all the relations ranked ahead the test relation are incorrect. Hits@3 is the proportion of samples whose ranks are no larger than three, and the definition of Hits@10 is similar. Generally, a better relation predictor achieves a lower mean rank. The higher the value of Hits@3, the better prediction performance. The maximum element value of this metric is 1.00. The results were compared with those of the following methods: TransE (Bordes *et al.*, 2013), TransH (Wang *et al.*, 2014) , Bilinear model (Jenatton *et al.*, 2012), and NTN (Socher *et al.*, 2013). The score functions of all the compared methods are listed in Table 2.

Table 2 Score functions of the compared methods

Method	Score function $S(\boldsymbol{e}_1, \boldsymbol{r}, \boldsymbol{e}_2)$
TransE	$\|\|\boldsymbol{e}_1 + \boldsymbol{r} - \boldsymbol{e}_2\|\|_2$
TransH	$\|\|\boldsymbol{e}_1 - \boldsymbol{w}_{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{e}_1\boldsymbol{w}_{\boldsymbol{r}} + \boldsymbol{d}_{\boldsymbol{r}} - (\boldsymbol{e}_2 - \boldsymbol{w}_{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{e}_2\boldsymbol{w}_{\boldsymbol{r}})\|\|_2$
Bilinear model	$\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{r}}\boldsymbol{e}_2$
NTN	$\boldsymbol{u}^{\mathrm{T}}f(\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{W}_{\boldsymbol{r}}\boldsymbol{e}_2 + \boldsymbol{V}_{r1}\boldsymbol{e}_1 + \boldsymbol{V}_{\boldsymbol{r}_2}\boldsymbol{e}_2 + \boldsymbol{b}_{\boldsymbol{r}})$
Joint embedding	$\boldsymbol{e}_1^{\mathrm{T}}\mathrm{diag}(\boldsymbol{r})\boldsymbol{e}_2$

3.3 Settings

In the experiments of our scheme, the embedding vectors of entities were constrained as unit vectors, while the relation embedding vectors were not restricted. Dimension k of the embedding vectors was set to 100 as default. Hyperparameters C's were set according to the validation sets. After the parameter investigation experiments, we finally set $C = 0.05$ on the Wordnet dataset, $C = 0.01$ on the FB13 dataset, and $C = 0.02$ on the FB15k dataset.

The dimensions of embeddings of the competing methods were also set to 100, except the dimension of NTN on the FB15k dataset which was set to 50 due to computation complexity. The learning rates and optimization steps were tuned based on the valid sets. The weights of the regular terms in NTN (namely hyperparameter λ) were fixed to 0.0001 as mentioned in Socher *et al.* (2013).

3.4 Relation embedding distribution

This section discusses the distribution of relation embedding vectors of our framework. A good embedding model embeds semantically similar relations into close vectors in the relation embedding space. Here, we present the visualization results and several examples of the neighborhood of relations to support our statement.

Fig. 2 shows the sketch maps of the distribution of the relation embedding spaces of our joint embedding model and TransE on the FB15k dataset by t-SNE (Maaten and Hinton, 2008). We classified the relations into 59 categories based on the relation name path and painted relations from the same category in the same color. Only the categories including the top 20 relations are shown in Fig. 2. The sketch maps indicate that the distribution of the relation embedding space of our joint embedding model is inner-class-concentrated.

Table 3 shows several relations and their three nearest neighbors in the relation embedding space. The relations are all selected from the FB15k dataset. Each relation is presented in a three-level index form as in the Freebase dataset originally. The nearest neighbors to the center relation embedded by our model are mainly synonyms or semantically close relations. In Table 3, it is clear that our method embeds semantically similar relations nearby in the relation embedding space, while TransE does not.

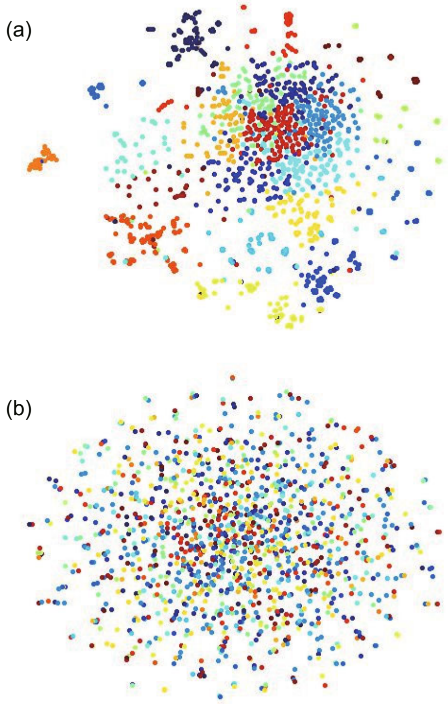

Fig. 2 Visualization of the distribution of the relation embedding spaces of our joint embedding model (a) and TransE (b) on the datasets of FB15k. Different colors represent different categories. References to color refer to the online version of this figure

The comparison with TransE suggests the advantage of our embedding model in indicating the semantic structure. TransE models the entities and relations in the same embedding space by satisfying $\boldsymbol{e}_1 + \boldsymbol{r} \approx \boldsymbol{e}_2$. Compared with it, our model embeds the relations into a different space from the entity spaces in which the relations are embedded more flexibly, encouraging the embedding result to satisfy the semantic rules of the relations.

Table 3 Some examples of the relation embedding distribution

Relation	Center relation
1	/film/film/film_art_direction_by
2	/olympics/olympic_medal/medal_winners
3	/award/award_category/category_of
Relation	**Cost-sensitive model (ours)**
1	/film/film_art_director/films_art_directed /film/film_set_designer/film_sets_designed /film/film/film_set_decoration_by
2	/olympics/olympic_medal_honor/medal /olympics/olympic_medal_honor/medalist /olympics/olympic_athlete/medals_won
3	/award/award/category /award/award_presenting_ /organization/award_categories_presented /award/award_category/presenting_organization
Relation	**TransE**
1	/people/profession/people_with_this_profession /education/education/educational_institution /location/hud_county_place/place
2	/location/hud_county_place/place /education/educational_institution/campuses /location/us_county/county_seat
3	/people/profession/people_with_this_profession /film/film_job/films_with_this_crew_job /common/webpage/topic

The table shows the three nearest relations to the center relation in the embedding spaces

3.5 Relation prediction

The results of all the compared methods on all the datasets are reported in Table 4. The results of 'joint embedding without sensitive learning' refer to the solution to problem (3), and results of 'joint embedding with sensitive learning' are the final results of our model. The results shown in boldface are the best ones. On all the three datasets, our joint

Table 4 Performances of the relation prediction evaluated by metrics MeanRank and Hits@3 (or Hits@10 on the FB15k dataset) using different models

Model	MeanRank			Hits@3 (Hits@10 on FB15k)		
	WordNet	FB13	FB15k	WordNet	FB13	FB15k
TransE	4.11	6.74	219.8	0.45	0.24	0.28
TransH	3.53	5.57	154.6	0.5	0.23	0.47
Bilinear	3.44	3.8	138.6	0.76	0.52	0.73
NTN	4.16	4.04	195.9	0.42	0.38	0.26
Joint embedding without cost-sensitive learning (ours)	3.47	4.83	166.5	0.59	0.43	0.40
Joint embedding with cost-sensitive learning (ours)	**1.58**	**2.99**	**25.5**	**0.95**	**0.69**	**0.79**

The results shown in boldface are the best ones model generally performs best, indicating that our scheme is effective on the relation prediction. The comparison between cost-sensitive learning and non-cost-sensitive learning embeddings indicates that the corresponding matrix $\boldsymbol{W}$ is helpful for the embedding. The performances on MeanRank indicate that the average performance of our prediction model is the best, while the high rate of Hits@3 (or Hits@10) indicates that our model is able to precisely predict the majority of the relations.

4 Conclusions

In this paper, we proposed a cost-sensitive learning-based joint embedding model to solve the knowledge embedding problem of embedding the entities and relations simultaneously. Based on the proposed model, we learned semantically meaningful embedding vectors for both entities and relations. In the model, a vector-product score function was defined to connect embedding vectors from different embedding spaces, and a weight matrix was involved to measure the relationships among the relations. This is helpful to take advantages of the higher-order context information in the knowledge base. Experimental results showed that the relation embedding distribution in the embedding space is semantically quite meaningful, and our embedding approach significantly improves the performances of the state-of-the-art methods on relation prediction.

References

Bollacker, K., Evans, C., Paritosh, P., *et al.*, 2008. Freebase: a collaboratively created graph database for structuring human knowledge. Proc. Int. Conf. on ACM SIGMOD Management of Data, p.1247-1250. https://doi.org/10.1145/1376616.1376746

Bordes, A., Weston, J., Collobert, R., *et al.*, 2011. Learning structured embeddings of knowledge bases. *Proc. Comput. Sci.*, **108**:345-354. https://doi.org/10.1016/j.procs.2017.05.045

Bordes, A., Usunier, N., Garcia-Duran, A., *et al.*, 2013. Translating embeddings for modeling multi-relational data. NIPS, p.2787-2795.

Bordes, A., Glorot, X., Weston, J., *et al.*, 2014. A semantic matching energy function for learning with multi-relational data. *Mach. Learn.*, **94**(2):233-259. https://doi.org/10.1007/s10994-013-5363-6

Chang, K.W., Yih, W.T., Meek, C., 2013. Multi-relational latent semantic analysis. Proc. Conf. on Empirical Methods in Natural Language Processing, p.1602-1612.

Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, **12**:2121-2159.

Elisseeff, A., Weston, J., 2001. A kernel method for multi-labelled classification. NIPS, p.681-687.

García-Durán, A., Bordes, A., Usunier, N., 2015. Composing Relationships with Translations. Technical Report No. hal-01167811, CNRS-Heudiasyc, Compiègne.

Getoor, L., Mihalkova, L., 2011. Learning statistical models from relational data. Proc. Int. Conf. on ACM SIGMOD Management of Data, p.1195-1198. https://doi.org/10.1145/1989323.1989451

Hoffmann, R., Zhang, C., Ling, X., *et al.*, 2011. Knowledge-based weak supervision for information extraction of overlapping relations. Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, p.541-550.

Jenatton, R., Roux, N.L., Bordes, A., *et al.*, 2012. A latent factor model for highly multi-relational data. NIPS, p.3167-3175.

Lin, Y.K., Liu, Z.Y., Luan, H.B., *et al.*, 2015. Modeling relation paths for representation learning of knowledge bases. arXiv: 1506.00379. http://arxiv.org/abs/1506.00379

Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**:2579-2605.

Miller, G.A., 1995. WordNet: a lexical database for English. *Commun. ACM*, **38**(11):39-41. https://doi.org/10.1145/219717.219748

Nickel, M., Tresp, V., Kriegel, H.P., 2011. A three-way model for collective learning on multi-relational data. Proc. 28th Int. Conf. on Machine Learning, p.809-816.

Singhal, A., 2012. Introducing the Knowledge Graph: Things, not Strings. Google. https://21stcenturylibrary.com/2012/05/21/introducing-googles-knowledge-graph-things-not-strings [Accessed on May 21, 2012].

Socher, R., Chen, D., Manning, C.D., *et al.*, 2013. Reasoning with neural tensor networks for knowledge base completion. NIPS, p.926-934.

Wang, Z., Zhang, J.W., Feng, J.L., *et al.*, 2014. Knowledge graph embedding by translating on hyperplanes. Proc. 28th Conf. on Artificial Intelligence, p.1112-1119.

Waters, R., 2012. Google to unveil search results overhaul. *Financial Times*, May.