Representation Learning with Entity Topics for Knowledge Graphs

Xin Ouyang[1(✉)], Yan Yang[1(✉)], Liang He[1,2], Qin Chen[1], and Jiacheng Zhang[1]

[1] Institute of Computer Applications, East China Normal University, Shanghai, China
{xouyang,yyang,lhe9191,qchen,jchzhang}@ica.stc.sh.cn
[2] Shanghai Engineering Research Center of Intelligent Service Robot, Shanghai, China

Abstract. Knowledge representation learning which represents triples as semantic embeddings has achieved tremendous success these years. Recent work aims at integrating the information of triples with texts, which has shown great advantages in alleviating the data sparsity problem. However, most of these methods are based on word-level information such as co-occurrence in texts, while ignoring the latent semantics of entities. In this paper, we propose an entity topic based representation learning (ETRL) method, which enhances the triple representations with the entity topics learned by the topic model. We evaluate our proposed method knowledge graph completion task. The experimental results show that our method outperforms most state-of-the-art methods. Specifically, we achieve a maximum improvement of 7.9% in terms of hits@10.

Keywords: Knowledge representation · Entity topics · Topic model · Knowledge graph completion

1 Introduction

Knowledge Graphs (KGs) aim at storing the facts of the real world with (h, r, t) triples, where h (or t) denotes the head (or tail) entity and r represents the relation between entities. Large-scale KGs such as Freebase [1] have been widely used in NLP tasks, including Question Answering (QA) [7]. However, with the increasing of triple amounts, KGs often suffer from data sparseness and incompleteness problems [11].

To alleviate these problems, Knowledge Representation Learning (KRL) which learns low-dimensional semantic embeddings of entities and relations has attracted extensive attention recently. Most of existing KRL methods learn the semantic representations from triples and show good performance in knowledge graph completion task. However, these methods cannot represent very well triples that suffer from sparsity. For example, more than 20% entities in FB15k [2]

G. Li et al. (Eds.): KSEM 2017, LNAI 10412, pp. 534–542, 2017.
DOI: 10.1007/978-3-319-63558-3_45

are associated with less than 20 triples, which leads to poor entity representations and diminishes the KRL performance. Therefore, recent researches turn to utilize the text information to enhance the performance of KRL [10,11]. In particular, Xie et al. [11] proposed DKRL model, which explored convolutional neural networks (CNN) for encoding entity descriptions. In [10], a TEKE model was proposed to learn the co-occurrence word networks for entities based on the Wikipedia pages. However, these methods only utilize the word-level information, and cannot well capture the latent topics of entities which are assumed to be useful to improve the KRL performance in this paper. For instance, "Agatha" is a playwright as demonstrated in the triple and has the descriptions as shown in Fig. 1. If a new entity "Joseph" has a description, namely "Joseph was an American satirical **novelist**, **short story writer** and **dramatist**", it is reasonable to infer that "Joseph" is also a playwright since he has the common topics like "writer" with "Agatha" and "dramatist" with "Playwright".

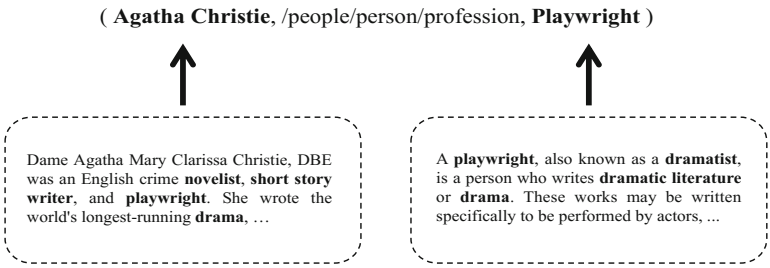

Fig. 1. Example of triple and entity descriptions.

Motivated by the potential of the latent topics for KRL, we propose an entity topic based representation learning (ETRL) method in this paper. Specifically, we first use entity descriptions from KGs for topic modeling, which shows strong semantic correlations with entities. The Nonnegative Matrix Factorization (NMF) model is applied to learn the topic representations of entities. We propose our basic model (ETRL(basic)) that treats the obtained topic representations as constants, and integrates them with the triple representations, and then propose the advanced model (ETRL(adv)) that learns both topic and triple representations simultaneously. The contributions of our works are as follows:

1. We propose two variants of ETRL models, which integrates the topic information based on the entity descriptions with the triple structure for knowledge representation learning.
2. We utilize projection matrices in ETRL to map topic and triple representations into a united semantic space.
3. We evaluate our models on knowledge graph completion task. The experimental results show that our models outperforms most of the state-of-the-art models.

The outline of this paper is as follows. Section 2 elaborates the recent works related to our works. Section 3 reveals the details of our method. Section 4 evaluates our models and analyzes the results. Section 5 concludes our works.

2 Related Works

In this section, we will introduce several works aiming at embedding the entities and relations into low-dimensional continuous vector space to address the shortages of large-scale KGs mentioned above.

TransE [2], one of the most famous methods, has shown its advantages among the methods proposed before. For each correct (h, r, t) triple, TransE regards that $h + r \approx t$, which means that the entity h is translated to the entity t through the relation r. TransE defines the score function as shown in (1).

$$e(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2 \tag{1}$$

However, it has issues in modeling the 1-N, N-1 and N-N relations [10]. There are various methods proposed to address these issues and achieved better performance than TransE, such as TransH [9], TransR [6], and KB2E [4].

The methods based on triples are suffering from data sparsity problem. There are increasing methods fusing external text information and triple information to alleviate this problem and enrich the semantic information of representations. The model proposed by Zhong et al. [12] jointly learns the triple information and text information from entity descriptions in word-level. DKRL [11] learns text representations with deep learning method CNN. TEKE [10] memorizes the words co-occurring in contexts of entities from wikipedia pages and then constructs co-occurrence networks to bridge the KG and text-corpus. However, most of these methods disregard the entity topics which contain strong semantic relevance between entities. Although the state-of-the-art method TEKE [10] has achieved some success in discovering entity topics for regarding co-occurrence words as topics of entities, it is still limited by the performance of entity linking tools and the completeness of external text-corpus.

3 Methodology

In this section, we will present our ETRL method. The structure of our method is shown in Fig. 2. Firstly, we use entity descriptions in KGs as our text-corpus and learn entity topic representations with NMF [5]. Then, we construct ETRL(basic) model to incorporate topic embeddings into triple embeddings. In addition, we build ETRL(adv) model to jointly learn two embeddings simultaneously. Finally, we train our models with Adagrad method.

3.1 Notations and Definitions

For a given KG, We denote E as the set of entities and R as the set of relations. Meanwhile, $\mathbf{E}$, $\mathbf{R}$ is the set of embeddings of E, R respectively. Then, we denote

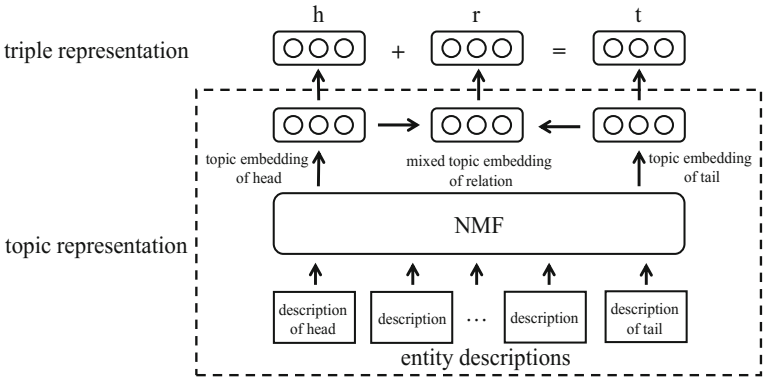

Fig. 2. The structure of ETRL(basic and adv).

T as the set of $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ triples in KG where $\mathbf{h}, \mathbf{t} \in \mathbf{E}$ and $\mathbf{r} \in \mathbf{R}$. T is also the train set for our models. For a given text-corpus D, we denote W as the set of words arising in D and denote $\mathbf{v_e}$ as the topic embedding of entity $e \in E$. It is noteworthy that all embeddings $\mathbf{h}, \mathbf{r}, \mathbf{t}, \mathbf{v_h}, \mathbf{v_t} \in \mathbb{R}^k$, where k is the embedding vector size.

3.2 Topic Representation of Entities

We use entity descriptions as our text-corpus for it is much easier for us to obtain entity descriptions, and most entities in KGs are associated with their own descriptions which can be regarded as semantic supplements.

Considering about the performance, efficiency, and complexity of topic models [8], we adopt NMF [5] to learn topic representations of entities. Regarding each entity description as input document, NMF is defined as follows:

$$M = VS \tag{2}$$

where M is the $n \times m$ word-frequency matrix of entities preprocessed from entity descriptions, V is the matrix of entity topic representations while S is the matrix of word topic representations.

We select Euclidean distance defined in (3) as the convergence criterion of NMF to factorize matrix M into entity and word topic embeddings.

$$L_{nmf} = \sum_{i=1}^{n} \sum_{j=1}^{m} \| M_{i,j} - \mathbf{v_{e_i}} \mathbf{s_{w_j}}^\top \|_2^2 \tag{3}$$

where $M_{i,j}$ is the frequency of word w_j appearing in the description of entity e_i, $\mathbf{v_{e_i}}$ and $\mathbf{s_{w_j}}$ are the topic embeddings of entity e_i and word w_j respectively. Note that in NMF, for all $e_i \in E$ and $w_j \in W$ we have $\mathbf{v_{e_i}} \geq 0$ and $\mathbf{s_{w_j}} \geq 0$. We pre-train the NMF and obtain the entity topic representations in store for triple representation learning.

3.3 Joint Representation Learning of Triples and Topics

The most important step is to map both entity topic and triple representations into same semantic vector space. Therefore, we propose ETRL(basic) which integrates two representations in training while treats the topic representations as constants. Then, we propose ETRL(adv) which learns the representations from both topic and triple information inspired by DKRL [11].

We present projection matrices to the alignment of topic and triple representations. Besides, these matrices can help models learn useful topic information and discard useless one automatically in training.

For head and tail, we directly project entity topic embeddings into the space of triples as follows:

$$\mathbf{h}_\perp = \mathbf{v_h} M_e + \mathbf{h} \tag{4}$$

$$\mathbf{t}_\perp = \mathbf{v_t} M_e + \mathbf{t} \tag{5}$$

where $\mathbf{v_h}$ (or $\mathbf{v_t}$) is the entity topic embedding of head (or tail), $\mathbf{h}$ and $\mathbf{t}$ are the original entity representations in (1), and M_e is the $k \times k$ projection matrix.

To strengthen the capability of representing complex relations, we average the head and tail topic representations as relation topic representations $\mathbf{v_r}$. Then we define $\mathbf{r}_\perp$ as follows:

$$\mathbf{r}_\perp = \mathbf{v_r} M_r + \mathbf{r} \tag{6}$$

where M_r is the $k \times k$ projection matrix. It is worth noting that, for complex relations, different head-tail pairs may have different $\mathbf{v_r}$ to affect $\mathbf{r}_\perp$, thus improving the variousness and robustness of relation representations.

Then we replace $\mathbf{h}, \mathbf{r}, \mathbf{t}$ in (1) with $\mathbf{h}_\perp, \mathbf{r}_\perp, \mathbf{t}_\perp$ and redefine the score function for both two models as follows:

$$e_\perp(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \|\mathbf{h}_\perp + \mathbf{r}_\perp - \mathbf{t}_\perp\|_2^2 \tag{7}$$

3.4 Loss Optimization and Training

ETRL(basic). To force the scores of correct triples as close as possible to zero and make the wrong's as far as possible, we define the loss function in (8).

$$L_s = \sum_{(h,r,t)\in T} \sum_{(h',r',t')\in T'} max(\gamma + e_\perp(\mathbf{h}, \mathbf{r}, \mathbf{t}) - e_\perp(\mathbf{h}', \mathbf{r}', \mathbf{t}'), 0) \tag{8}$$

where $max(\cdot, 0)$ is hinge loss, T is the train set consisting correct triples while T' is the set of negative samples denoted in (9).

$$T' = \{(h', r, t)|h' \in E\} \cup \{(h, r', t)|r' \in R\} \cup \{(h, r, t')|t' \in E\} \tag{9}$$

where h', r' and t' are selected by a certain probability to satisfy $\forall(\mathbf{h}, \mathbf{r}, \mathbf{t}) \in T'$, $(\mathbf{h}, \mathbf{r}, \mathbf{t}) \notin T$. We follow the'bern' method mentioned in TransH [9] to select the negative samples.

ETRL(adv). Inspired by the works of DKRL [11], we propose an advanced model to jointly learn both the topic and triple representations. The loss function is defined in (10).

$$L_a = L_s + L_{nmf} \tag{10}$$

where we optimize triple loss (8) and topic loss (3) simultaneously.

Training. We adopt Adagrad in the optimization for a better performance than SGD. In order to accelerate the speed of convergence, we initialize the entities and relations with the vectors produced by TransE and initialize the projection matrices as identify matrices.

4 Experiments

In this section, we will evaluate our models on knowledge graph completion task. Then we analyze results comparing with other state-of-the-art methods.

4.1 Datasets and Experiment Settings

Datasets. In this paper, we adopt FB15k dataset extracted from Freebase [1] as experimental dataset. FB15k is first proposed by the authors of TransE [2] and is widely used in KRL methods. The statistics of FB15k are listed in Table 1.

Table 1. Statistics of dataset.

Dataset	Entity	Relation	Train	Valid	Test
FB15k	14,951	1,345	483,142	50,000	59,071

Text-Corpus. We extract the English entity descriptions from the latest Freebase Dump [3] in FB15k as our text-corpus. The average number of words in each entity description is 124. Then we filter out the stop words, connect phrases which contain the entity names and remove the suffixes of the words in text-corpus.

Experiment Settings. We train our two models and set the best hyperparameters as follows: the number of iterations $iter = 1000$, embeddings size and topic numbers $k = 100$, margin $\gamma = 2$ and the learning rate scaling factor $\epsilon = 1 \times 10^{-7}$ by experience. We select TransE and two state-of-the-art methods, TEKE and DKRL as our baseline. We also compare our models to other methods such as TransH and TransR.

4.2 Knowledge Graph Completion

The task of knowledge graph completion (also called link prediction) aims at completing the triples which have missing components. For instance, given a

triple (h, r, t) where head entity h is missing, the model firstly fill all entities $e \in E$ as candidate entities in the triple and then compute the score of triple (e, r, t) for each entity. Finally, it ranks the scores of candidate entities and completes the triple with the entity e which has the lowest score.

Two evaluation metrics [2] are adopted in this task: mean rank (MR) and hits@10. The former means the average rank of the correct one in the list of candidates while the latter means the rate of correct one ranked in the top 10. We also follow the two settings as "Raw" and "Filter" mentioned in TransE [2].

Entity Prediction Results. We remove the head and tail entities respectively of triples in test set and compute MR and hits@10. Then we take the average MR of heads and tails as the final results. The results listed in Table 2 show that our models especially ETRL(adv) outperform all the baseline except the mean rank(filter). By comparing with other experimental results listed in Table 2 we find that:

1. All evaluation metrics of our models are better than baselines which shows that the rich semantic information in entity topics can improve the performance. Furthermore, our models could be treated as the topic supplement for TransE and achieve 28.6% improvement than TransE in terms of hits@10.
2. Because TransR has more projection matrices [6], mean rank(filter) of TransR is better than our models. This shows that it is necessary to adopt more projection matrices for the learning of entities and relations.
3. The performance of ETRL(adv) is better than ETRL(basic). We can infer that joint models are more suitable for incorporating topic and triple embeddings with more smoothy influence.

Table 2. Entity prediction results on FB15k.

Metrics	Mean Rank		Hits@10	
	Raw	Filter	Raw	Filter
TransE	243	125	34.9	47.1
TransH	212	87	45.7	64.4
TransR	198	**77**	48.2	68.7
DKRL(CNN)	200	113	44.3	57.6
DKRL(CNN) + TransE	181	91	49.6	67.4
TEKE_TransE	233	79	43.5	67.6
ETRL(basic)	174	90	54.8	74.8
ETRL(adv)	**170**	83	**60.3**	**75.5**

Complex Relation Modeling Problem. Complex relation modeling problem, which leads to low performance in modeling 1-N, N-1 and N-N relations, is one of the most important problems in KRL. According to the unbalance of

the ratio of head to tail, researchers divide relations into 1-1, 1-N, N-1 and N-N relations where the latter three are called complex relations. We test the ability of modeling complex relations of our models. The results listed in Table 3 show our models outperform others in modeling 1-N, N-1 and N-N relations. Based on the results, we believe that with the addition of projection matrices both on entities and relations, our models can have advantages in handling the complex relation modeling problem.

Table 3. Entity prediction for complex relation problem on FB15k.

Metrics	Hits@10			
	1-1	1-N	N-1	N-N
TransE	43.7	42.7	42.5	48.6
TransH	66.2	63.7	56.0	65.9
TransR	**79.0**	63.3	62.3	70.7
TEKE_TransE	47.6	61.2	63.8	76.5
ETRL(basic)	63.1	**66.5**	63.1	77.4
ETRL(adv)	62.5	65.0	**67.1**	**78.8**

5 Conclusion and Future Work

In this paper, to make full use of entity topic information in entity descriptions, we propose ETRL model using NMF to learn entity topic embeddings and mapping both topic and triple embeddings into same semantic space with projection matrices. Experimental results show that our models outperform most state-of-the-art methods which enhancing KRL with external text information. However, there also are some deficiencies to be solved in our future works. To lower the error brought by entity topic information, we will extend our models with more projection matrices to overcome the complex relations problem inspired by TransR [6]. We will extend our models to other state-of-the-art triple based representation models and try to improve the performance of them.

Acknowledgments. This work was supported by the National Key Technology Support Program (No. 2015BAH01F02), Shanghai Municipal Commission of Economy and Information Under Grant Project (No. 201602024), the Natural Science Foundation of Shanghai (No. 17ZR1444900)and the Science and Technology Commission of Shanghai Municipality (No. 15PJ1401700).

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June, pp. 1247–1250 (2008)

2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS, pp. 2787–2795 (2013)
3. Google: Freebase data dumps. https://developers.google.com/freebase/data
4. He, S., Liu, K., Ji, G., Zhao, J.: Learning to represent knowledge graphs with gaussian embedding. In: CIKM, pp. 623–632. ACM (2015)
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
6. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: AAAI, pp. 2181–2187 (2015)
7. Miller, A.H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: EMNLP (2016)
8. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: EMNLP-CoNLL, pp. 952–961 (2012)
9. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI, pp. 1112–1119. Citeseer (2014)
10. Wang, Z., Li, J.: Text-enhanced representation learning for knowledge graph. In: IJCAI, pp. 1293–1299. AAAI Press (2016)
11. Xie, R., Liu, Z., Jia, J., Luan, H., Sun, M.: Representation learning of knowledge graphs with entity descriptions. In: AAAI, pp. 2659–2665 (2016)
12. Zhong, H., Zhang, J., Wang, Z., Wan, H., Chen, Z.: Aligning knowledge and text embeddings by entity descriptions. In: EMNLP, pp. 267–272 (2015)