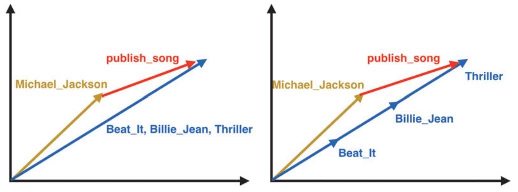
Proceedings, Fifteenth International Conference on
Principles of Knowledge Representation and Reasoning (KR 2016)

Knowledge Graph Embedding by Flexible Translation

Jun Feng[1], Minlie Huang[1], Mingdong Wang[2], Mantong Zhou[1], Yu Hao[1], Xiaoyan Zhu[1]
[1]State Key Lab. of Intelligent Technology and Systems, National Lab. for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, PR China
[2]Dept. of Physics, Tsinghua University

Abstract

Knowledge graph embedding refers to projecting entities and relations in knowledge graph into continuous vector spaces. Current state-of-the-art models are translation-based model, which build embeddings by treating relation as translation from head entity to tail entity. However, previous models is too strict to model the complex and diverse entities and relations(e.g. symmetric/transitive/one-to-many/many-to-many relations). To address these issues, we propose a new principle to allow flexible translation between entity and relation vectors. We can design a novel score function to favor flexible translation for each translation-based models without increasing model complexity. To evaluate the proposed principle, we incorporate it into previous method and conduct triple classification on benchmark datasets. Experimental results show that the principle can remarkably improve the performance compared with several state-of-the-art baselines.



(a) TransE (b) Flexible Translation

Figure 1: Illustration of TransE and our proposed Flexible Translation. There are three triples, which share the same head entity ("Michael_Jackson") and the same relation ("publish_song"), while having three different tail entities ("Beat_it", "Billie_Jean", and "Thriller"). (a) TransE can hardly distinguish different tail entities as they all approximated to the sum of head vector and relation vector. (b) Instead of strictly constraining **h+r=t**, our principle is to enforce that **h+r** has the same direction with **t**.

Introduction

Knowledge graphs such as Wordnet and Freebase, which are representations of multi-relational data, have become very important resources to support many AI related tasks, such as natural language understanding, question answering, web search, etc. A knowledge graph is usually represented by a directed graph, in which nodes refer to entities and edges refer to relations between entities, or simply by a set of triples *head entity, relation, tail entity* ((h, r, t) for short). Although there have been substantial achievements in building large-scale knowledge graph, the general paradigm to support computing is not clear. Indeed, traditional knowledge graphs are symbolic and logical frameworks which are not flexible enough to be fruitfully exported, especially to statistical learning approaches which require the knowledge to be computable in numerical forms.

Recently, knowledge graph embedding, which projects entities or/and relations between entities into a continuous vector space, has been a new proposal to offer the powerful capability of computing on knowledge graphs. In this paradigm, the embedding representation of a single entity/relation encodes the global information of the entire knowledge graph, since the representation is obtained by minimizing a global loss function involving all entities and relations. Furthermore, by concerning knowledge computation, the embedding representations are beneficial to a variety of applications such as question answering and web search. Taking knowledge graph completion as an example, we can simply judge the correctness of a triple (h, r, t) by checking the compatibility score of the embedding vectors of $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$.

A variety of approaches have been explored for knowledge graph embedding, such as general linear based models (Bordes et al. 2011), bilinear based models (Jenatton et al. 2012; Sutskever, Tenenbaum, and Salakhutdinov 2009), neural network based models (Socher et al. 2013), and translation based models (Bordes et al. 2013). However, there are limitations of the existing methods. For example, general linear based models can hardly capture the correlations between entities and relations. Bilinear based models can only model linear interactions and is not able to approach more complex scoring functions. In addition, the complexity of neural network based models is too high to handle large-scaled knowledge bases.

Another line of previous models are translation based models, including TransE (Bordes et al. 2013), TransH (Wang et al. 2014), TransR (Lin et al. 2015) and PTransE (Lin, Liu, and Sun 2015), which are efficient and achieve the state-of-art performance. All these models adopt the same principle, $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$, if (h,r,t) is a fact of knowledge base. More specifically, current translation based methods use the same form of distance measure $f_r(h,t) = \parallel \mathbf{h}_r + \mathbf{r} - \mathbf{t}_r \parallel_{l_{1/2}}$ (Manhattan/Euclidean distance) to quantify the compatibility score of a triple, where $\mathbf{h}_r$ and $\mathbf{t}_r$ are the embedding vectors of head and tail entities which are projected into the relation-specific space.

However, the principle $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$ is too strict to model the complex and diverse objects including entities and relations(e.g. symmetric/transitive/one-many/many-to-one/many-many relations)in knowledge graph. Taking a one-to-many relation *publish_song* as the example, we have triples such as (*Michael_Jackson, publish_song, Beat_It*), (*Michael_Jackson, publish_song, Billie_Jean*) and (*Michael_Jackson, publish_song, Thriller*). As shown in Figure 1(a), considering the ideal embedding where $\mathbf{h}+\mathbf{r}=\mathbf{t}$, the entities *Beat_It*, *Billie_Jean* and *Thriller* will get the same embedding vectors.

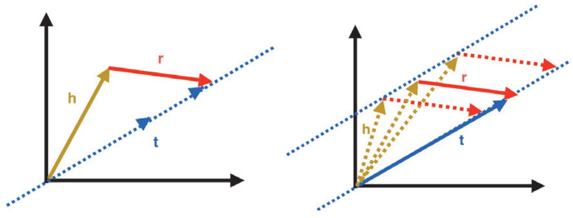
To address the above issues, we propose a new principle to allow *Flexible Translation*(FT) to model complex and diverse entities and relations. The central idea of FT is well-motivated (see Fig. 1): instead of strictly enforcing $\mathbf{h}+\mathbf{r}=\mathbf{t}$, we only constrain that the direction of $\mathbf{h}+\mathbf{r}$ (or $\mathbf{t}-\mathbf{r}$) is the same as that of $\mathbf{t}$ (or $\mathbf{h}$), but allow flexible magnitude of resulting vectors. Therefore, unlike previous method assumes $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$ if (h,r,t) holds, FT takes $\mathbf{h}+\mathbf{r} \approx \alpha\mathbf{t}$(or $\mathbf{t}-\mathbf{r} \approx \alpha\mathbf{h}$), where the flexibility is reflected in α. Thus we design a new function to score the compatibility of a triple by the inner product between the sum of head entity vector and relation vector **h + r** and tail vector $\mathbf{t}$ instead of using the Manhattan/Euclidean distance which are commonly used in previous models. This new principle can better capture the complexity and diversity of knowledge graphs. As previously presented in Figure 1(b), the ideal embedding where $\mathbf{h}+\mathbf{r} \approx \alpha\mathbf{t}$ is that the embedding vectors of three entities *Beat_It*, *Billie_Jean* and *Thriller* are with the same direction but are of different vector magnitudes.

More importantly, the proposed principle is quite general and efficient as it can be adopted to translation-based models, such as TransE, TransH, and TransR. Experiments show that when applying the principle to these models, remarkable improvements can be obtained over the original models, and notably, there is no increase in model complexity.

Related Work

There are a variety of models for knowledge graph embedding. We survey translation-based models which are the mainstream models for knowledge graph embedding, and other related models.

Translation-based models build embeddings by treating relation as translation from head entity to tail entity. Though these models differ in score functions, they all share quite similar principle $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$. TransE (Bordes et al. 2013) represents relationships by translation vectors in an embedding space. TransH (Wang et al. 2014) projects entity embeddings $\mathbf{h}$ and $\mathbf{t}$ to the hyperplane and applies the same assumption as in TransE. TransR (Lin et al. 2015) addresses the issue that some entities are similar in the entity space but comparably different in other specific aspects. PTransE considers the relation path while training embeddings. SSE (Guo et al. 2015) incorporates geometrically based regularization terms, constructed by using additional semantic categories of entities.



(a) (b)

Figure 2: The different embedding vector range of tail entity($\mathbf{t}$) and head entity($\mathbf{h}$). Given h and r, the range of $\mathbf{t}$ is a line; given t and r, the range of $\mathbf{h}$ is a plane.

There are many other models proposed for knowledge graph embedding. We also use Structured Embedding (Bordes et al. 2011), Latent Factor Model (Jenatton et al. 2012; Sutskever, Tenenbaum, and Salakhutdinov 2009), Neural Tensor Network (Socher et al. 2013), Semantic Matching Energy (Bordes et al. 2014) and Single Layer Model (Socher et al. 2013) as our baselines in the experiments.

Method

To address the issues existing in previous translation based methods as mentioned previously, we propose a general principle named *Flexible Translation*(FT). We can better model the complex and diverse entities/relations with the idea of flexible translation.

Let's introduce some notations. S denotes a set of golden triples, while S' denotes a set of corrupted triples. A triple (h,r,t) consists of two entities $h,t \in E$ (the set of entities) and relation $r \in R$ (the set of relations). We use the bold letters $\mathbf{h},\mathbf{r}$ and $\mathbf{t}$ to denote the corresponding embedding vectors.

Flexible Translation

Let's start with the limitations of previous translation based models which adopt an over-strict principle: $\mathbf{h}+\mathbf{r} \approx \mathbf{t}$.More specifically, if with the ideal embedding using the function $\mathbf{h}+\mathbf{r}=\mathbf{t}$ when (h,r,t) holds, we can obtain: 1) if (h_1,r,t_1), (t_1,r,t_2) and $(h_1,r,t_2)$are both correct, r is a transitive relation. In addition, we obtain two conflicting equations: $\mathbf{h}_1+\mathbf{r}=\mathbf{t}_2, \mathbf{h}_1+2\mathbf{r}=\mathbf{t}_2$; 2) if a set of triples $(h,r,t_i), \forall i \in 0,\cdots,n$ hold, r is a one-to-many relation and $\mathbf{t}_0=\cdots=\mathbf{t}_n$; 3) if (h,r_1,t), (h,r_2,t) and (h,r_3,t) hold, we get $\mathbf{r}_1=\mathbf{r}_2=\mathbf{r}_3$.

To alleviate the problem of previous translation based methods and maintain the high efficiency, we apply $\mathbf{h}+\mathbf{r} \approx \alpha\mathbf{t}, \alpha > 0$ when (h,r,t) holds. That means we only need to maintain the directions of vectors $\mathbf{h}+\mathbf{r}$ and $\mathbf{t}$, but ignore their magnitudes. Therefore, 1) if r is a transitive relation, we get $\mathbf{h_1}+\mathbf{r}=\alpha_1\mathbf{t_1}$, $\mathbf{t}_1+\mathbf{r}=\alpha_2\mathbf{t}_2$ and $\mathbf{h_1}+\mathbf{r}=\alpha_1\mathbf{t_2}$; 2) if r is a one-to-many relation, we get $\mathbf{t}_0 = \frac{\mathbf{h+r}}{\alpha_0}, \cdots, \mathbf{t_n} = \frac{\mathbf{h+r}}{\alpha_n}$; 3) if (h, r_1, t), (h, r_2, t) and (h, r_3, t) hold, $\mathbf{r}_1 = \alpha_1\mathbf{t} - \mathbf{h}$, $\mathbf{r}_2 = \alpha_2\mathbf{t} - \mathbf{h}$ and $\mathbf{r}_3 = \alpha_3\mathbf{t} - \mathbf{h}$.

The score function is then defined as follows:

$$f_r(h,t) = (\mathbf{h} + \mathbf{r})^\top \mathbf{t} \tag{1}$$

However, with the score function, the constraints on head entity(h) and tail entity(t) are unbalanced. More specifically, under the constraints, the range of $\mathbf{h}$ is a line, and the range of $\mathbf{t}$ is a plane. Considering the perfect no-error embedding, we discuss the constraints on tail entity and head entity separately. As shown in Figure 2(a), when the embedding vectors $\mathbf{h}$ and $\mathbf{r}$ hold, the range of the no-error embedding vector $\mathbf{t}$ is a vector with the right arrow on the dotted line. However, as shown in Figure 2(b), when the embedding vectors $\mathbf{r}$ and $\mathbf{t}$ are known, the range of the perfect embedding vector $\mathbf{h}$ is a vector with the starting point on the lower dotted line and the ending point on the upper dotted line.

Since head entity and tail entity need to have the same effect during training, the constraints on them should be balanced. To this end, we design a Flexible Translation to address the unbalanced constraint problem. We modify the score function as follows:

$$f_r(h,t) = (\mathbf{h} + \mathbf{r})^\top \mathbf{t} + \mathbf{h}^\top(\mathbf{t} - \mathbf{r}) \tag{2}$$

In FT, the score is expected to be higher for a golden triple while lower for a corrupted one.

Connection to other Models

To show the generality of our principle, we show here how other translation based models can be improved using the idea of Flexible Translation. Unlike previous models which apply $\mathbf{h}_r + \mathbf{r} \approx \mathbf{t}_r$ if (h, r, t) is a golden triple, we adopts $\mathbf{h}_r + \mathbf{r} \approx \alpha\mathbf{t}_r$. Accordingly, we define the score function $f_r(h,t) = (\mathbf{h}_r + \mathbf{r})^\top \mathbf{t}_r + \mathbf{h}_r^\top(\mathbf{t}_r - \mathbf{r})$, where $\mathbf{h}_r$ and $\mathbf{t}_r$ are the embedding vectors of head and tail entities which projected into the relation-specific space.

In TransE, the entity and relation embedding vectors are in the same space, say $\mathbf{h}_r = \mathbf{h}, \mathbf{t}_r = \mathbf{t}$. The improved model of TransE is called TransE-FT. As the improved model follows the principle of Flexible Translation, it can fix the flaws of TransE when dealing with one-to-many/many-to-one/many-to-many relations.

In TransH, entity embedding vectors are projected into a relation-specific hyperplane $\mathbf{w}_r$, say $\mathbf{h}_r = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r, \mathbf{t}_r = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r$. We call the enhanced model as TransH-FT.

In TransR models, $\mathbf{h}_r = \mathbf{h}\mathbf{M}_r, \mathbf{t}_r = \mathbf{t}\mathbf{M}_r$, where entities are projected from the entity space to the relation space by $\mathbf{M}_r$. The improved model is called TransR-FT.

Training Objective

All models are trained with contrastive max-margin objective functions. The objective is to ensure that a triple $(h, r, t) \in S$ in the golden set should have a higher score than a triple $(h', r, t') \in S'$ in the corrupted triple set, as follows:

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'} max(0, \gamma - f_r(\mathbf{h}, \mathbf{t}) + f_r(\mathbf{h}', \mathbf{t}')) \tag{3}$$

where $\gamma > 0$ is a margin hyperparameter. S is the training set of golden triples. S' is the set of corrupted triples. The corrupted triples is generated from the training triples with either the head or tail entity replaced by a random entity (but not both at the same time). We adopt the mini-batched stochastic gradient descent(SGD) to optimize the objective function.

Table 1: The statistics of the corpora.

Dataset	#R.	#Ent.	#Train	#Valid	#Test
WN18	18	40,943	141,442	5,000	5,000
FB15K	1,345	14,951	483,142	50,000	59,071
WN11	11	38,696	112,581	2,609	10,544
FB13	13	75,043	316,232	5,908	23,733

Table 2: Comparison of accuracy on triple classification(%)

DataSets	WN11	FB13	FB15K
SE (Bordes et al. 2011)	53.0	75.2	-
SME (Bordes et al. 2014)	70.0	63.7	-
SLM (Socher et al. 2013)	69.9	85.3	-
LFM (Jenatton et al. 2012)	73.8	84.3	-
NTN (Socher et al. 2013)	70.4	**87.1**	68.5
TransE (Bordes et al. 2013)	75.9	70.9	79.6
TransE-FT(ours)	**86.4**	**82.1**	**90.5**
TransH (Wang et al. 2014)	77.7	76.5	79.0
TransH-FT(ours)	**78.3**	**80.7**	**82.1**
TransR (Lin et al. 2015)	85.5	74.7	81.7
TransR-FT(ours)	**86.6**	**82.9**	**88.9**

Complexity Analysis

To analyze the efficiency of our model, we compare the number of parameters between different models. TransE-FT, TransH-FT and TransR-FT maintain the same number of parameters as their original models TransE, TransH and TransR, respectively. Therefore, the proposed principle is able to improve the previous translation based models without sacrificing the efficiency.

Experiments

To justify our proposed principle, we apply the general principle to several models including TransE, TransH , and TransR, and conduct extensive experiments to compare the enhanced variant models (TransE-FT, TransH-FT and TransR- FT) with the original ones. First, we evaluate our models on triple classification (Socher et al. 2013)

Three benchmark datasets are tested in the experiments: WN18 (Bordes et al. 2013) and WN11 (Socher et al. 2013) which is extracted from Wordnet (Miller 1995); and two dense subgraphs of Freebase (Bollacker 2008), FB15K (Bordes et al. 2013) and FB13 (Socher et al. 2013). Table 1 shows the statistics of these data sets.

Triple Classification

Similar to the experiment in (Bordes et al. 2013; Wang et al. 2014; Lin et al. 2015), we evaluate our model on triple classification. Triple classification is a binary classification task which predict whether a given triple (h, r, t) is correct or not. This task is applied for answering question such as *Does Michael Jackson publish the song Beat it?*.We use three data sets in this task: WN11 and FB13 released in NTN (Socher et al. 2013); FB15K used in TransR (Lin et al. 2015).

Evaluation protocol. Following the protocol in NTN (Socher et al. 2013), we set a relation-specific threshold T_r for prediction and then, for a triple (h, r, t), if the similarity score obtained by f_r is above T_r, the triple (h, r, t) is predicted as positive, otherwise negative. The relation-specific threshold T_r is determined by maximizing the classification accuracy on a validation set.

Implementation. We compare our models with the baseline methods reported in (Lin et al. 2015) for WN11, FB13 and FB15K.

For training TransE-FT, the optimal configuration are: $\lambda = 0.01, \gamma = 2.25, k = 100, B = 960$ on WN11; $\lambda = 0.005, \gamma = 2, k = 100, B = 960$ on FB13; $\lambda = 0.002, \gamma = 0.5, k = 100, B = 960$ on FB15K. To training TransH-FT, the optimal configuration are: $\lambda = 0.05, \gamma = 1.5, k = 100, B = 960$ on WN11; $\lambda = 0.005, \gamma = 1.5, k = 50, B = 960$ on FB13; $\lambda = 0.01, \gamma = 0.5, k = 100, B = 960$ on FB15K. For experiments with TransR-FT, the best configurations are: $\lambda = 0.0001, \gamma = 2.5, k = 50, B = 960$ on WN11; $\lambda = 0.001, \gamma = 2.5, k = 50, B = 960$ on FB13; $\lambda = 0.001, \gamma = 0.1, d = 50, B = 120$ on FB15K. The number of training epochs is limited to $1,000$ for TransE-FT and 500 for TransH-FT and TransR-FT.

Experiment Result. Evaluation results are reported in Tabel 2. It demonstrates that our TransE-FT model outperforms all the baseline models significantly including TransE, TransH, and even TransR on WN11 and FB15K. This result shows that, with the help of our Flexible Translation, TransE can get better result than the its modified models including TransH and TransR with the same model complexity. On FB13, TransE-FT beats all baseline models except the NTN model. As described in (Wang et al. 2014; Lin et al. 2015), FB13 is much denser than WN11 and FB15K where strong correlations exist between entities, and NTN can achieve better results by learning complicated correlations using tensor transformation from dense graph of FB13. It also shows that the enhancement models including TransE-FT, TransH-FT and TransR-FT consistently outperform their original models on all datasets. This observation demonstrates the superiority and generality of our Flexible Translation idea.

Conclusion

In this paper, we propose a new principle *Flexible Translation*(FT) for knowledge graph embedding. The central idea of FT is to ensure the direction of vectors during translation and to allow flexible magnitude of targeting vectors. The proposed principle can better capture the complex and diverse relations and entities (e.g. symmetric/transitive/one-many/many-to-one/many-many relations) in knowledge graphs. In addition, the principle of flexible translation is quite general and can be adopted to other embedding models without increasing model complexity. We conduct extensive experiment on benchmark datasets for the task triple classification, and results show that following our principle, TransE, TransH, and TransR can be substantially improved.

Acknowledgments

This work was partly supported by the National Basic Research Program (973 Program) under grant No.2012CB316301/2013CB329403, the National Science Foundation of China under grant No.61272227/61332007, and the Beijing Higher Education Young Elite Teacher Project.

References

Bordes, A.; Weston, J.; Collobert, R.; Bengio, Y.; et al. 2011. Learning structured embeddings of knowledge bases. In *AAAI*.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*.

Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning*.

Guo, S.; Wang, Q.; Wang, B.; Wang, L.; and Guo, L. 2015. Semantically smooth knowledge graph embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Jenatton, R.; Roux, N. L.; Bordes, A.; and Obozinski, G. R. 2012. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems*.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion.

Lin, Y.; Liu, Z.; and Sun, M. 2015. Modeling relation paths for representation learning of knowledge bases. *arXiv preprint arXiv:1506.00379*.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*.

Sutskever, I.; Tenenbaum, J. B.; and Salakhutdinov, R. 2009. Modelling relational data using bayesian clustered tensor factorization. In *Advances in neural information processing systems*.

Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.