TransA: An Adaptive Approach for Knowledge Graph Embedding

Han Xiao\textsuperscript{1}, Minlie Huang\textsuperscript{1}, Hao Yu\textsuperscript{1}, Xiaoyan Zhu\textsuperscript{1}

\textsuperscript{1}Department of Computer Science and Technology, State Key Lab on Intelligent Technology and Systems, National Lab for Information Science and Technology, Tsinghua University, Beijing, China

Abstract

Knowledge representation is a major topic in AI, and many studies attempt to represent entities and relations of knowledge base in a continuous vector space. Among these attempts, translation-based methods build entity and relation vectors by minimizing the translation loss from a head entity to a tail one. In spite of the success of these methods, translation-based methods also suffer from the oversimplified loss metric, and are not competitive enough to model various and complex entities/relations in knowledge bases. To address this issue, we propose TransA, an adaptive metric approach for embedding, utilizing the metric learning ideas to provide a more flexible embedding method. Experiments are conducted on the benchmark datasets and our proposed method makes significant and consistent improvements over the state-of-the-art baselines.

Introduction

Knowledge graphs such as Wordnet (Miller 1995) and Freebase (Bollacker et al. 2008) play an important role in AI researches and applications. Recent researches such as query expansion prefer involving knowledge graphs (Bao et al. 2014) while some industrial applications such as question answering robots are also powered by knowledge graphs (Fader, Zettlemoyer, and Etzioni 2014). However, knowledge graphs are symbolic and logical, where numerical machine learning methods could hardly be applied. This disadvantage is one of the most important challenges for the usage of knowledge graph. To provide a general paradigm to support computing on knowledge graph, various knowledge graph embedding methods have been proposed, such as TransE (Bordes et al. 2013), TransH (Wang et al. 2014) and TransR (Lin et al. 2015).

Embedding is a novel approach to address the representation and reasoning problem for knowledge graph. It transforms entities and relations into continuous vector spaces, where knowledge graph completion and knowledge classification can be done. Most commonly, knowledge graph is composed by triples \((h, r, t)\) where a head entity \(h\), a relation \(r\) and a tail entity \(t\) are presented. Among all the proposed embedding approaches, geometry-based methods are an important branch, yielding the state-of-the-art predictive performance. More specifically, geometry-based embedding methods represent an entity or a relation as \(k\)-dimensional vector, then define a score function \(f_s(h, t)\) to measure the plausibility of a triple \((h, r, t)\). Such approaches almost follow the same geometric principle \(h + r \approx t\) and apply the same loss metric \(||h + r - t||^2\) but differ in the relation space where a head entity \(h\) connects to a tail entity \(t\).

However, the loss metric in translation-based models is oversimplified. This flaw makes the current embedding methods incompetent to model various and complex entities/relations in knowledge base.

Firstly, due to the inflexibility of loss metric, current translation-based methods apply spherical equipotential hyper-surfaces with different plausibilities, where more near to the centre, more plausible the triple is. As illustrated in Fig.1, spherical equipotential hyper-surfaces are applied in (a), so it is difficult to identify the matched tail entities from the unmatched ones. As a common sense in knowledge graph, complex relations, such as one-to-many, many-to-one and many-to-many relations, always lead to complex embedding topologies. Though complex embedding situation is
We classify prior studies into two lines: one is the translation-based embedding methods and the other includes many other embedding methods.

Translation-Based Embedding Methods
All the translation-based methods share a common principle $h + r = t$, but differ in defining the relation-related space where a head entity $h$ connects to a tail entity $t$. This principle indicates that $t$ should be the nearest neighbour of $(h + r)$. Hence, the translation-based methods all have the same form of score function that applies Euclidean distance to measure the loss, as follows:

$$f_r(h, t) = \| h_r + r_r - t_r \|^2$$

where $h_r, t_r$ are the entity embedding vectors projected in the relation-specific space. Note that this branch of methods keeps the state-of-the-art performance.

- **TransE** (Bordes et al. 2013) lays the entities in the original space, say $h_r = h, t_r = t$.
- **TransH** (Wang et al. 2014) projects the entities into a hyperplane for addressing the issue of complex relation embedding, say $h_r = h - w_r h w_r, t_r = t - w_r t w_r$.
- **TransR** (Lin et al. 2015) transforms the entities by the same matrix to also address the issue of complex relation embedding, as: $h_r = M_r h, t_r = M_r t$.

Projecting entities into different hyperplanes or transforming entities by different matrices allow entities to play different roles under different embedding situations. However, as the “Introduction” argues, these methods are incompetent to model complex knowledge graphs well and particularly perform unsatisfactorily in various and complex entities/relations situation, because of the oversimplified metric.

**TransM** (Fan et al. 2014) pre-calculates the distinct weight for each training triple to perform better.

Other Embedding Methods
There are also many other models for knowledge graph embedding.

**Unstructured Model (UM).** The UM (Bordes et al. 2012) is a simplified version of TransE by setting all the relation vectors to zero $r = 0$. Obviously, relation is not considered in this model.

**Structured Embedding (SE).** The SE model (Bordes et al. 2011) applies two relation-related matrices, one for head and the other for tail. The score function is defined as $f_r(h, t) = \| M_{r, h} h - M_{r, t} t \|^2$. According to (Socher et al. 2013), this model cannot capture the relationships among entities and relations.

**Single Layer Model (SLM).** SLM applies neural network to knowledge graph embedding. The score function is defined as

$$f_r(h, t) = u_r^\top g(M_{r,1} h + M_{r,2} t)$$

Note that SLM is a special case of NTN when the zero tensors are applied. (Collobert and Weston 2008) had proposed...
a similar method but applied this approach into the language model.

**Semantic Matching Energy (SME).** The SME model (Bordes et al. 2012) (Bordes et al. 2014) attempts to capture the correlations between entities and relations by matrix product and Hadamard product. The score functions are defined as follows:

\[
\begin{align*}
    f_r &= (M_1 h + M_2 r + b_3)\top (M_3 t + M_4 r + b_2) \\
    f_r &= (M_1 h \otimes M_2 r + b_3)\top (M_3 t \otimes M_4 r + b_2)
\end{align*}
\]

where \(M_1, M_2, M_3\) and \(M_4\) are weight matrices, \(\otimes\) is the Hadamard product, \(b_3\) and \(b_2\) are bias vectors. In some recent work (Bordes et al. 2014), the second form of score function is re-defined with 3-way tensors instead of matrices.

**Latent Factor Model (LFM).** The LFM (Jenatton et al. 2012) uses the second-order correlations between entities by a quadratic form, defined as \(f_r(h, t) = h\top W_r t\).

**Neural Tensor Network (NTN).** The NTN model (Socher et al. 2013) defines an expressive score function for non-negative weight matrix that corresponds to the adaptive metric. Different from the traditional score functions, we take the absolute value, since we want to measure the absolute loss between \((h + r)\) and \(t\). Furthermore, we would list two main reasons for the applied absolute operator.

On one hand, the absolute operator makes the score function as a well-defined norm only under the condition that all the entries of \(W_r\) are non-negative. A well-defined norm is necessary for most metric learning scenes (Kulis 2012), and the non-negative condition could be achieved more easily than PSD, so it generalises the common metric learning algebraic form for better rendering the knowledge topologies. Expanding our score function as an induced norm \(N_r(e) = \sqrt{f_r(h, t)}\) where \(e = h + r - t\). Obviously, \(N_r\) is non-negative, identical and absolute homogeneous. Besides with the easy-to-verified inequality \(N_r(e_1 + e_2) = \sqrt{|e_1| + |e_2|}|W_r|e_1 + e_2| \leq \sqrt{|e_1||W_r|e_1| + |e_2||W_r|e_2|} = N_r(e_1) + N_r(e_2)\), the triangle inequality is hold. Totally, absolute operators make the metric a norm with an easy-to-achieve condition, helping to generalise the representation ability.

On the other hand, in geometry, negative or positive values indicate the downward or upward direction, while in our approach, we do not consider this factor. Let’s see an instance as shown in Fig 2. For the entity Goniolf, the x-axis component of its loss vector is negative, thus enlarging this component would make the overall loss smaller, while this case is supposed to make the overall loss larger. As a result, absolute operator is critical to our approach. For a numerical example without absolute operator, when the embedding dimension is two, weight matrix is \([0, 1; 1, 0]\) and the loss vector \((h + r - t) = (e_1, e_2)\), the overall loss would be \(2|e_1|e_2|\). If \(e_1 \geq 0\) and \(e_2 \leq 0\), much absolute larger \(e_2\) would reduce the overall loss and this is not desired.

**Perspective from Equipotential Surfaces**

TransA shares almost the same geometric explanations with other translation-based methods, but they differ in the loss metric. For other translation-based methods, the equipotential hyper-surfaces are spheres as the Euclidean distance defines:

\[
||(t - h) - r||^2 = C
\]

where \(C\) means the threshold or the equipotential value. However, for TransA, the equipotential hyper-surfaces are elliptical surfaces as the Mahalanobis distance of absolute loss states (Kulis 2012):

\[
|(t - h) - r\top W_r(t - h) - r| = C
\]

Note that the elliptical hyper-surfaces would be distorted a bit as the absolute operator applied, but this makes no difference for analysing the performance of TransA. As we know,
Taking other constraints into account, the target function can be defined as follows:

$$\min \sum_{(h,r,t) \in \Delta} \sum_{(h',r',t') \in \Delta'} [f_r(h,t) + \gamma - f_{r'}(h',t')]_+ +$$

$$\lambda \left( \sum_{r \in R} ||W_r||_F^2 \right) + C \left( \sum_{e \in E} ||e||_2^2 + \sum_{r \in R} ||r||_2^2 \right)$$

s.t. $$|W_r|_{ij} \geq 0$$

where $$[\cdot]_+ = \max(0, \cdot)$$. \(\Delta\) is the set of golden triples and \(\Delta'\) is the set of incorrect ones, \(\gamma\) is the margin that separates the positive and negative triples. \(||\cdot||_F\) is the F-norm of matrix, \(C\) controls the scaling degree, and \(\lambda\) controls the regularization of adaptive weight matrix. The \(E\) means the set of entities and the \(R\) means the set of relations. At each round of training process, \(W_r\) could be worked out directly by setting the derivation to zero. Then, in order to ensure the non-negative condition of \(W_r\), we set all the negative entries of \(W_r\) to zero.

$$W_r = -\sum_{(h,r,t) \in \Delta} \left( ||h + r - t||^T \right) + \sum_{(h',r',t') \in \Delta'} \left( ||h' + r' - t'||^T \right)$$

As to the complexity of our model, the weight matrix is completely calculated by the existing embedding vectors, which means TransA almost has the same free parameter number as TransE. As to the efficiency of our model, the weight matrix has a closed solution, which speeds up the training process to a large extent.

**Experiments**

We evaluate the proposed model on two benchmark tasks: link prediction and triples classification. Experiments are conducted on four public datasets that are the subsets of Wordnet and Freebase. The statistics of these datasets are listed in Tab[1].

ATPE is short for “Averaged Triple number Per Entity”. This quantity measures the diversity and complexity of datasets. Commonly, more triples lead to more complex structures of knowledge graph. To express the more complex structures, entities would be distributed variously and complexly. Overall, embedding methods produce less satisfactory results in the datasets with higher ATPE, because a large ATPE means a various and complex entities/relations embedding situation.

**Link Prediction**

Link prediction aims to predict a missing entity given the other entity and the relation. In this task, we predict \(t\) given \((h, r, *)\), or predict \(h\) given \((*, r, t)\). The WN18 and FB15K datasets are the benchmark datasets for this task.

**Evaluation Protocol.** We follow the same protocol as used in TransE (Bordes et al. 2013), TransH (Wang et al. 2014) and TransR (Lin et al. 2015). For each testing triple \((h, r, t)\), we replace the tail \(t\) by every entity \(e\) in the knowledge graph and calculate a dissimilarity score with the score...
2. FB15K is a very various and complex entities/relations embedding situation, because its ATPE is absolutely highest among all the datasets. However, TransA performs better than other baselines on this dataset, indicating that TransA performs better in various and complex entities/relations embedding situation. WN18 may be less complex than FB15K because of a smaller ATPE. Compared to TransE, the relative improvement of TransA on WN18 is 5.7% while that on FB15K is 95.2%. This comparison shows TransA has more advantages in the various and complex embedding environment.

3. TransA promotes the performance for 1-1 relations, which means TransA generally promotes the performance on simple relations. TransA also promotes the performance for 1-N, N-1, N-N relations which demonstrates TransA works better for complex relation embedding.

4. Compared to TransR, better performance of TransA means the feature weighting and the generalised metric form led by absolute operators, have significant benefits, as analysed.

5. Compared to Adaptive Metric (PSD) which applies the score function $f_r(h, t) = (h + r - t)^\top W_r(h + r - t)$ and constrains $W_r$ as PSD, TransA is more competent, because our score function with non-negative matrix condition and absolute operator produces a more flexible representation than that with PSD matrix condition does, as analysed in “Adaptive Metric Approach”.

6. TransA performs bad in Mean Rank on WN18 dataset. Digging into the detailed situation, we discover there are 27 testing triples (0.54% of the testing set) whose ranks are more than 30,000, and these few cases would make about 162 mean rank loss. The tail or head entity of all these triples have never been co-occurring with the corresponding relation in the training set. It is the insufficient training data that leads to the over-distorted weight matrix and the over-distorted weight matrix is responsible for the bad Mean Rank.

---

2. ATPE: Averaged Triple number Per Entity. Triples are summed up from all the #Train, #Valid and #Test.

3. Mapping properties of relations follow the same rules in (Bordes et al. 2013).
Table 3: Evaluation results on FB15K by mapping properties of relations(%) 

| Relation Category | Predicting Head(HITS@10) | Predicting Tail(HITS@10) |
|-------------------|---------------------------|---------------------------|
|                   | 1-1 | 1-N | N-1 | N-N | 1-1 | 1-N | N-1 | N-N |
| SE (Bordes et al. 2011) | 35.6 | 62.6 | 17.2 | 37.5 | 34.9 | 14.6 | 68.3 | 41.3 |
| SME (Bordes et al. 2012)  | 35.1 | 53.7 | 19.0 | 40.3 | 32.7 | 14.9 | 61.6 | 43.3 |
| TransE (Bordes et al. 2013) | 43.7 | 65.7 | 18.2 | 47.2 | 43.7 | 19.7 | 66.7 | 50.0 |
| TransH (Wang et al. 2014) | 66.8 | 87.6 | 28.7 | 64.5 | 65.5 | 39.8 | 83.3 | 67.2 |
| TransR (Lin et al. 2015)  | 78.8 | 89.2 | 34.1 | 69.2 | 79.2 | 37.4 | 90.4 | 72.1 |
| TransA               | 86.8 | 95.4 | 42.7 | 77.8 | 86.7 | 54.3 | 94.4 | 80.6 |

Table 4: Triples classification: accuracies(%) for different embedding methods 

| Methods                  | WN11 | FB13 | Avg. |
|--------------------------|------|------|------|
| LFM                      | 73.8 | 84.3 | 79.0 |
| NTN                      | 70.4 | 87.1 | 78.8 |
| TransE                   | 75.9 | 81.5 | 78.7 |
| TransH                   | 78.8 | 83.3 | 81.1 |
| TransR                   | 85.9 | 82.5 | 84.2 |
| Adaptive Metric (PSD)    | 81.4 | 87.1 | 84.3 |
| TransA                   | 83.2 | 87.3 | 85.3 |

Triples Classification

Triples classification is a classical task in knowledge base embedding, which aims at predicting whether a given triple \((h, r, t)\) is correct or not. Our evaluation protocol is the same as prior studies. Besides, WN11 and FB13 are the benchmark datasets for this task. Evaluation of classification needs negative labels. The datasets have already been built with negative triples, where each correct triple is corrupted to get one negative triple.

Evaluation Protocol. The decision rule is as follows: for a triple \((h, r, t)\), if \(f_r(h, t)\) is below a threshold \(\sigma_r\), then positive; otherwise negative. The thresholds \(\{\sigma_r\}\) are determined on the validation dataset. The final accuracy is based on how many triples are classified correctly.

Implementation. As all methods use the same datasets, we directly copy the results of different methods from the literature. We have tried several settings on the validation dataset to get the best configuration for both Adaptive Metric (PSD) and TransA. The optimal configurations are: “bern” sampling, \(\alpha = 0.02\), \(k = 50\), \(\gamma = 10.0\), \(C = 0.2\) on WN11, and “bern” sampling, \(\alpha = 0.002\), \(k = 200\), \(\gamma = 3.0\), \(C = 0.00002\) on FB13.

Results. Accuracies are reported in Tab[4] and Fig[3] According to “Adaptive Metric Approach” section, we could work out the weights by \(LDL\ Decomposition\) for each relation. Because the minimal weight is too small to make a significant analysis, we choose the median one to represent relative small weight. Thus, “Weight Difference” is calculated by \(\frac{\text{MaximalWeight} - \text{MedianWeight}}{	ext{MedianWeight}}\). Bigger the weight difference is, more significant effect, the feature weighting makes. Notably, scaling by the median weight makes the weight differences comparable to each other. We observe that:

1. Overall, TransA yields the best average accuracy, illustrating the effectiveness of TransA.
2. Accuracies vary with the weight difference, meaning the feature weighting benefits the accuracies. This proves the theoretical analysis and the effectiveness of TransA.
3. Compared to Adaptive Metric (PSD), TransA performs better, because our score function with non-negative matrix condition and absolute operator leads to a more flexible representation than that with PSD matrix condition does.

Conclusion

In this paper, we propose TransA, a translation-based knowledge graph embedding method with an adaptive and flexible metric. TransA applies elliptical equipotential hypersurfaces to characterise the embedding topologies and weights several specific feature dimensions for a relation to avoid much noise. Thus, our adaptive metric approach could effectively model various and complex entities/relations in knowledge base. Experiments are conducted with two benchmark tasks and the results show TransA achieves consistent and significant improvements over the current state-of-the-art baselines. To reproduce our results, our codes and data will be published in github.
References

[Bao et al. 2014] Bao, J.; Duan, N.; Zhou, M.; and Zhao, T. 2014. Knowledge-based question answering as machine translation. Cell 2:6.

[Bollacker et al. 2008] Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 1247–1250. ACM.

[Bordes et al. 2011] Bordes, A.; Weston, J.; Collobert, R.; Bengio, Y.; et al. 2011. Learning structured embeddings of knowledge bases. In Proceedings of the Twenty-fifth AAAI Conference on Artificial Intelligence.

[Bordes et al. 2012] Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In International Conference on Artificial Intelligence and Statistics, 127–135.

[Bordes et al. 2013] 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, 2787–2795.

[Bordes et al. 2014] Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2014. A semantic matching energy function for learning with multi-relational data. Machine Learning 94(2):233–259.

[Collobert and Weston 2008] Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, 160–167. ACM.

[Fader, Zettlemoyer, and Etzioni 2014] Fader, A.; Zettlemoyer, L.; and Etzioni, O. 2014. Open question answering over curated and extracted knowledge bases. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1156–1165. ACM.

[Fan et al. 2014] Fan, M.; Zhou, Q.; Chang, E.; and Zheng, T. F. 2014. Transition-based knowledge graph embedding with relational mapping properties. In Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation, 328–337.

[Golub and Van Loan 2012] Golub, G. H., and Van Loan, C. F. 2012. Matrix computations, volume 3. JHU Press.

[Guo et al. 2015] Guo, S.; Wang, Q.; Wang, B.; Wang, L.; and Guo, L. 2015. Semantically smooth knowledge graph embedding. In Proceedings of ACL.

[Jenatton et al. 2012] 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, 3167–3175.

[Kulis 2012] Kulis, B. 2012. Metric learning: A survey. Foundations & Trends in Machine Learning 5(4):287–364.

[Lin et al. 2015] Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[Lin, Liu, and Sun 2015] Lin, Y.; Liu, Z.; and Sun, M. 2015. Modeling relation paths for representation learning of knowledge bases. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics.

[Miller 1995] Miller, G. A. 1995. Wordnet: a lexical database for english. Communications of the ACM 38(11):39–41.

[Nickel, Tresp, and Kriegel 2011] Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In Proceedings of the 28th international conference on machine learning (ICML-11), 809–816.

[Nickel, Tresp, and Kriegel 2012] Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2012. Factorizing yago: scalable machine learning for linked data. In Proceedings of the 21st international conference on World Wide Web, 271–280. ACM.

[Socher et al. 2013] 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, 926–934.

[Wang and Sun 2014] Wang, F., and Sun, J. 2014. Survey on distance metric learning and dimensionality reduction in data mining. Data Mining and Knowledge Discovery 1–31.

[Wang et al. 2014] 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, 1112–1119.

[Wang, Wang, and Guo 2015] Wang, Q.; Wang, B.; and Guo, L. 2015. Knowledge base completion using embeddings and rules. In Proceedings of the 24th International Joint Conference on Artificial Intelligence.