# Interaction Network Analysis Using Semantic Similarity Based on Translation Embeddings

Awais Manzoor Bajwa[1], Diego Collarana[2(✉)], and Maria-Esther Vidal[3,4,5]

[1] University of Bonn, Bonn, Germany
bajwaa@cs.uni-bonn.de
[2] Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin, Germany
collaran@cs.uni-bonn.de
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
maria.vidal@tib.eu
[4] L3S Research Centre, Leibniz University of Hannover, Hannover, Germany
[5] Universidad Simón Bolívar, Caracas, Venezuela

**Abstract.** Biomedical knowledge graphs such as STITCH, SIDER, and Drugbank provide the basis for the discovery of associations between biomedical entities, e.g., interactions between drugs and targets. Link prediction is a paramount task and represents a building block for supporting knowledge discovery. Although several approaches have been proposed for effectively predicting links, the role of semantics has not been studied in depth. In this work, we tackle the problem of discovering interactions between drugs and targets, and propose SimTransE, a machine learning-based approach that solves this problem effectively. SimTransE relies on translating embeddings to model drug-target interactions and values of similarity across them. Grounded on the vectorial representation of drug-target interactions, SimTransE is able to discover novel drug-target interactions. We empirically study SimTransE using state-of-the-art benchmarks and approaches. Experimental results suggest that SimTransE is competitive with the state of the art, representing, thus, an effective alternative for knowledge discovery in the biomedical domain.

**Keywords:** Knowledge graphs · Embeddings · Similarity function

## 1 Introduction

The discovery of interactions among entities is one of the main link prediction tasks over knowledge graphs. Specifically, the problem of drug-target interaction discovery, i.e., proteins that are targets of drugs, is a crucial task, given the fact, that on average, bringing a new drug to the market, costs ≈ $1.8 billion and takes more than 10 years. Several approaches have been defined to tackle the


M. Acosta et al. (Eds.): SEMANTiCS 2019, LNCS 11702, pp. 249–255, 2019.
https://doi.org/10.1007/978-3-030-33220-4_18

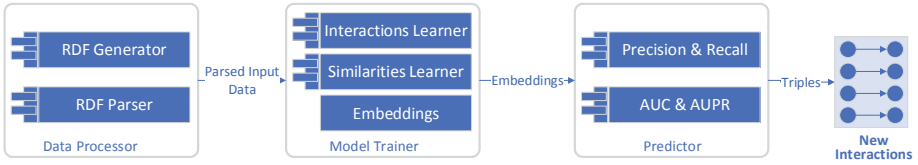



**Fig. 1. The Architecture.** SimTransE receives an RDF knowledge graph and similarities among its entities. The output is a set of predicted interactions.

problem of drug-target interaction discovery (e.g., [2,4]). Albeit effective, existing approaches are not able to exploit the semantics encoded in the main features of the drugs or targets to enhance prediction. We present SimTransE approach that exploits both similarities between entities, e.g., drugs and target, as well as their connections in a knowledge graph. These features are considered by SimTransE to represent entities into a vector space. SimTransE is based on TransE, which utilizes the gradient descent optimization method to learn the embeddings based on relations stated in a knowledge graphs. Similarly, SimTransE optimizes the distance between embeddings, considering the existing interactions between drugs and targets, but additionally, SimTransE takes into consideration domain similarity values between drugs and between targets. Embeddings generated by SimTransE are utilized to predict new interactions by applying the homophily principle[1]. We conduct an empirical evaluation to assess the quality of SimTransE with respect to TransE and a benchmarks of interactions between drugs and targets. Our observed results suggest that considering similarity empowers SimTransE and allows for the discovery of interactions between drugs and targets that could be identified by baseline version of TransE.

## 2 The SimTransE Approach

After reviewing different approaches such as [2,4], we realize the benefits that integrating the entity-entity similarity (e.g., target-target, drug-drug, and target-drug) into a learning model can bring. The intuition behind this work is that vector embedding-based approaches effectively combine different dimensions of the input data to learn embeddings. As a result, embeddings merge different dimensions of the data giving a multi-dimensional entity representation. We present SimTransE, an approach that maps each entity into multi-dimensional vector space considering entity-entity similarities to improve the results of the link prediction task. Thus, SimTransE is a vector embedding based machine learning model to learn a bipartite graph interactions and predict unknown interactions.

### 2.1 Architecture

The SimTransE architecture comprises a pipeline with three main components. Figure 1 shows the interaction between these components and the data flowing

[1] https://en.wikipedia.org/wiki/Homophily.

among them. The *Data Processor* receives an RDF graph and creates dictionaries and matrices understandable by SimTransE. Three sets of *entity dictionaries* are created, i.e., left entities (the subjects), right entities (the objects), and relational entities. These dictionaries are used throughout the pipeline to create vector embeddings. Secondly, two different sets of binary *sparse matrices* are created. One representing the **positive and negative interactions** of entities. Lastly, similarity matrices are built, i.e., given the **m** number of left entities and **n** number of right entities, we prepare two square matrices where the similarity score between entities from **m** to **n** are kept. The *Model Trainer* component receives as input the entity and interaction dictionaries and similarity matrices. The Model Trainer resorts to the **stochastic gradient descent** method to optimize the position and direction of the embeddings in a vector space. The *Model Trainer* uses interactions and similarities between entities to solve the optimization problem, and generates embeddings as output; (Table 1 shows the SimTransE interaction and objective functions). The *Predictor* component takes the generated embedding vectors, interactions, and thresholds. Using the embeddings and thresholds, this component iterates over all the entities and identifies interactions of each entity with every other entity. The *Predictor* component calculates the precision and recall. Additionally, the Area Under Receiver (AUC) and the Area Under the Precision-Recall Curve (AUPRC) are calculated.

### 2.2 Learning Vector Embeddings

State-of-art approaches use only connectivity patterns between entities to learn the embeddings and perform predictions. Using just interactions among entities is not enough real-world applications where domain-specific knowledge plays a relevant role (e.g., during the prediction of drug-target interactions [8]). There are very few known interactions and the ratio of positive to negative classes is large, impacting, this, in the accuracy of the predictions. To tackle the problem of unbalance ratio of positive to negative classes, SimTransE incorporates not only entities interactions but similarities between entities during the learning process. SimTransE creates duplicate positive classes and adds a set of positive examples, which are generated using the similarity matrices. The similarity score is considered as the weight of example in the learning process.

SimTransE[2] analyses the interactions and similarities between entities to learn the embeddings. SimTransE is based on the work "Translating Embeddings for Modelling Multi-relational Data" (TransE [1]). SimTransE intuition relies on the basic idea of TransE, i.e., if two entities interacts with each other, then the sum of first entity vector and relation vector should be approximately equal to the second vector. If there is no interaction between the two entities, the sum of first entity vector and relation vector will be far from the second entity vector. Using the same principle, SimTransE locates vectors using the similarities as well, and adds a new condition in the learning model that states similar entities

[2] Algorithms are documented in our repository https://github.com/RDF-Molecules/SimTransE.

**Table 1.** SimTransE interaction and objective function to learn embeddings

| Interaction functions | Objective functions |
|---|---|
| $\begin{cases} h + l \approx t, & \text{if h interacts (l) t} \\ h + l \not\approx t, & \text{otherwise} \end{cases}$ | $L_i = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} [\gamma + d(h+\ell,t) - d(h'+\ell,t')]_+$ |
| $\begin{cases} h1 + l \approx h2, & \text{if h1 similar h2} \\ h1 + l \not\approx h2, & \text{otherwise} \end{cases}$ | $L_s = \sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in SI_{(h,\ell,t)}} [d(h+\ell,t) - d(h'+\ell,t')]_+$ |

should be closer than the dissimilar ones. Interactions are generated based on the homophily principle that states that similar entities tend to interact with similar entity. Further, we rely on thresholds captured from the meaning of the similarity metrics and to decide when two entities can be considered similar. Then the **stochastic gradient descent** optimization method is performed; a mini-batch of drug-target interactions is generated according to a training set **S** of interactions. The embeddings are updated during the learning process with two objective functions: (1) $\boldsymbol{L_i}$ minimizes the distance whenever this is greater between actual and a corrupted triple with respect to the relation among them; and (2) $\boldsymbol{L_s}$ minimizes the distance according to the similarity between the actual and self-generated similar triples. The learning process stops when reaching the total number of epochs, or depending on a threshold about the distance between the generated embeddings and the training set.

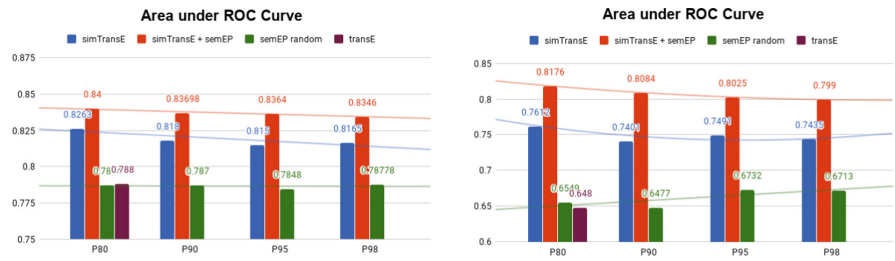
### 2.3 Predicting Links

The fundamental task of link prediction is to identify a relations between two entities. Yang et al. [9] define the link prediction formally as a task in a network $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is the set of nodes and $\mathbf{E}$ is the set of edges. The main challenge to be achieved in this task is to predict whether there is or will be a link $e(u, v)$ between a pair of nodes u and v $\in \mathbf{V}$ and $e(u, v) \notin \mathbf{E}$. To perform link prediction, SimTransE uses the trained vector embeddings and calculates the distance of each entity to every other entity with respect to the relation between them. Based on this calculated distance and a given threshold, SimTransE decides if the input entities are or not related. SimTransE ranks each entity on the basis of distance and assigns a probability by comparing it with the distance of other entities. If this probability is greater than the given threshold, then SimTransE considers the link in the output.

To evaluate link prediction we measure: **Precision**, the ratio of correctly predicted interactions to total predictions; **Recall**, the ratio of correctly predicted interactions to expect predictions; **Area under Precision-Recall Curve** , we calculate the area under precision recall curve as the metric to evaluate our model, it does not consider true negatives since neither of both precision and recall consider true negatives; Finally, we measure the **Area under ROC Curve**, to evaluate our method since it works best when the problem of imbalanced classes exist in the dataset [9] (Fig. 2).

## 3 Empirical Evaluation

We empirically study the effectiveness of SimTransE on the problem of predicting links. We assess the following research questions: **(RQ1)** Is SimTransE able to perform as good as the state-of-the-art similarity measures? **(RQ2)** Does SimTransE perform well on the task of link prediction when applied to data with lots of connections? To answer these questions, we evaluate SimTransE on a state-of-the-art benchmark of drug-target interactions [8]; we report only on the results of interactions between drugs and targets of the type Nuclear Receptor and Ion Channel. TransE [1] is the baseline of the experiment. Additionally, we utilize the link prediction technique, SemEP [4], that extracts interaction from highly connected partitions of a knowledge graph; these interactions are utilized to enhance the set of input interactions. Furthermore, we compute the drug-drug and target-target similarity matrices; drug similarities are computed using SIMCOMP [3] while target similarities are computed using a normalised Smith-Waterman score [5].




(a) AUC results on Ion Channel (b) AUC results on Nuclear Receptor

**Fig. 2.** SimTransE exhibits good performance in both datasets.

**Results and Discussion:** From the output of SimTransE, we calculated: *true and false positives* and *true and false negatives*. From these values, we derived *Precision*, *Recall*, *AUC*, and *AUPR*[3]. We apply a **blocking** method on the generated similarity-based interactions, through percentiles, i.e., four percentiles are considered: 80, 90, 95, and 100. Link prediction is validated following 10-fold cross-validation, and we report the mean across the results of the ten folds. Based on the observed outcomes, we can positively answer **RQ1**, i.e., SimTransE performs well on all the datasets, and outperforms the baseline method TransE in all cases. These results suggest that similarities between entities, e.g., drugs and targets, have a positive impact on both the learning process and the link prediction tasks. We observe, as well, that by increasing the number of connections between drugs and target (e.g., by using SemEP results) the effectiveness

[3] Source code and formulas to calculate Precision, Recall, AUC, and AUPR are documented in our repository https://github.com/RDF-Molecules/SimTransE.

of the approach improve even further. Few interactions are not predicted properly although they are present in the training set. For most of them, we find that drugs and targets with few numbers of interactions are difficult to train for SimTransE. This situation is improved after using the interactions predicted from SemEP. Therefore, **RQ2** is positively answered too.

## 4 Conclusions

In this paper, we presented SimTransE, a method to analyze interactions in knowledge graphs to predict links, based on the vectorization of the entities. To learn the embeddings, SimTransE uses not only the interactions among entities but also values of similarity between them. To test the accuracy of SimTransE, we compared its results against TransE, a prediction model for translational embeddings that uses only interactions among entities. SimTransE exhibited high accuracy and competitive result and outperformed TransE, one of the state-of-the-art approaches. The observed results suggest that combining interaction and similarity related semantics in the embeddings empowers the prediction model over knowledge graphs. In future work, we plan to conduct a more exhaustive evaluation to guarantee the reproducibility of the results, as well as the comparison with other embedding creation models, e.g., TransH [6] and TransG [7].

## References

1. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: 27th Annual Conference on Neural Information Processing Systems, pp. 2787–2795. Nevada, US (2013)
2. Fakhraei, S., Huang, B., Raschid, L., Getoor, L.: Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(5), 775–787 (2014)
3. Hattori, M., Okuno, Y., Goto, S., Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J. Am. Chem. Soc. **125**(39), 11853–11865 (2003)
4. Palma, G., Vidal, M.-E., Raschid, L.: Drug-target interaction prediction using semantic similarity and edge partitioning. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 131–146. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_9
5. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. **147**(1), 195–197 (1981)
6. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, pp. 1112–1119. Québec City, Québec, Canada (2014)
7. Xiao, H., Huang, M., Zhu, X.: Transg : a generative model for knowledge graph embedding. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, vol. 1: Long Papers (2016)

8. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Supplement of paper: prediction of drug-target interaction networks from the integration of chemical and genomic spaces (2018). Last accessed 15 Apr. 2018
9. Yang, Y., Lichtenwalter, R.N., Chawla, N.V.: Evaluating link prediction methods. Knowl. Inf. Syst. **45**(3), 751–782 (2015)