Knowledge Graph Representation Learning Based Drug Informatics

Rajeev Verma
Dept. of Electrical Engineering
Indian Institute of Technology Patna
rajeev.ee15@iitp.ac.in

Dr. Preetam Kumar
Dept. of Electrical Engineering
Indian Institute of Technology Patna
pkumar@iitp.ac.in

***Abstract*—Drug Discovery and assessment is increasingly a complex albeit an acute process. Current approaches to the study of drugs involving clinical evaluation and post-marketing surveillance is very time consuming and costly. Though the machine-learning based systems have been proposed in the literature, but those systems are task-dependent. This work presents a rather data-intensive approach to the study of drugs using the large-scale DrugBank dataset, a Linked Open Dataset to further the process of drug discovery. We used Representation Learning approach on the large-scale drug dataset which not only provides automated framework for feature learning for machine learning task but also helps in prediction of new facts in the knowledge base through Link-Prediction. The proposed representation learning system is promising to be used general approach to solve different problems and we experiment on Drug-Drug Interaction prediction and Drug-Target Prediction with comparable with respect to other work in the former task.**

***Index Terms*—Drug Discovery, Bioinformatics, Knowledge Graph, Machine Learning, Relational Learning, Deep Learning**

I. Introduction

The modern data-intensive approach to the bioinformatics is possible due to the widely managed Knowledge databases. These Knowledge Bases are available as Linked Open Data(LOD), a graph-structured data model also known as Knowledge Graph. Knowledge Graphs are at the centre of current state-of-the-art in Question-Answering, Information Retrieval, social-network analysis,etc.

A Knowledge Graph is a multi-relational data consisting of the triples of form *(head,relation,tail)* where *head* and *tail* are entities(nodes) and the *relation* denotes how the two entities are related(directed edge). The fragment of a typical Knowledge graph is shown in figure 1. The one of the triples shown is *(Sujoy Ghosh, director_of, Kahaani)* representing the very fact. A consistently large amount of effort is required to maintain these Knowledge Graphs. Even the largest of these like DBPedia, Wikidata, Freebase, etc. suffer from data-incompleteness and missing information affects the related downstream task. Relational Learning focuses on developing efficient tools to automatically add new facts(predict missing information), without requiring extra knowledge. The fundamental task in Relational Learning is Link Prediction(predicting new triples or predicting the correctness of unseen triples) and the graph-structure in which the knowledge is encoded supposedly

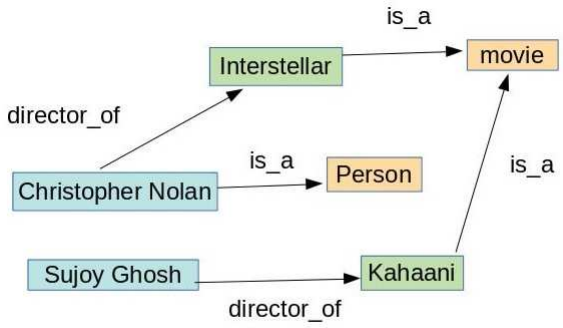

Fig. 1. A typical Knowledge Graph structure: The Entities are in the boxes with a labeled edge relating the two entities.

contains a lot-more information through the neighborhood and locality structure. The presence of triples *(Christopher Nolan, director_of, Interstellar)* and *(Christopher Nolan,is_a, Person)* and the triple *(Sujoy Ghosh, director_of, Kahaani)* would imply that *(Sujoy Ghosh, is_a, Person)* is also true. Thus, by harnessing such neighborhood structure would help us in the completion of Knowledge Graphs and predicting new information in the Knowledge bases. This kind of modeling has shown great strength in areas as Recommender systems, Google Search etc. and is currently the actively researched field in Machine Learning and Deep Learning community.

Representation Learning: The performance of any Machine Learning system depends on data-representation. Unsupervised Feature Learning or Representation Learning as it is known, is the major advantage in modern Deep Learning Systems. Representation Learning in Knowledge Graph aims to embed the components of knowledge Graphs into some *d*-dimensional continuous vector spaces, preserving the structure and the properties of Knowledge Graph.

We extend this approach to Bioinformatics to the actively maintained Knowledge Base DrugBank.

II. Related Works

A. Representation Learning In Knowledge Graphs

Representation Learning in Knowledge Graphs methods' embed entities as *d*-dimensional vectors, known as *embeddings* and relation as operators that combine the vectors

of two entities(*head* and *tail* of triple), or the correctness of a triple is scored by the relational scoring function. As per [1], the two major techniques for Knowledge Graph Representation Learning are *translational distance models* and *semantic matching models*. Translational Distance Models employ distance-based scoring functions while the Semantic Matching Models employ similarity-based scoring functions. The major works falling in former category are TransE [2] and the other related models TransR [3], TransD [4]. The works in latter category are RESCAL [5], DistMult [6], Neural Tensor Network(NTN) [7]. Other Neural Network based Link-Predictor models are ConvE [8], RDF2vec [9] which is a Random Walk based method inspired from works in Graph Representation Learning: DeepWalk [10], Node2vec [11].

B. Machine Learning works in Drugs' Study

Other Existing works [12] [13] [14] [15] on Drugs' study using machine learning are based on similarity criteria among drugs based on properties as drug-targets, drug-proteins, side-effects. Note that these methods are application dependent like predicting drug-drug interactions, predicting drug-target pairs etc. and no general framework to the knowledge has been proposed which models the considered problems together.

III. Knowledge Graph Description

For our study, we used Drugbank v5.0 [16] database. We used available Linked Open Data, Bio2RDF [17]. Bio2RDF is an open-source project that creates a large RDF graph interlinking data from many biological databases related to biological entities such as drugs,proteins,genes etc. The Dataset statistics is given in Table I.

TABLE I
Dataset Statistics

property	**count**
ENTITIES	521,612
TRIPLES	3,149,168
RELATIONS	104
DRUGS	8,097

The common relations are *drug-classification category*, *calculated-properties*, *ddi-interactor-in*, *ingredient*, *target*, *enzyme*, *mechanism-of-action*, *protein-binding*, *toxicity*, *gene-sequence*, etc. A graph-structure of the knowledge graph is shown in Figure 2.

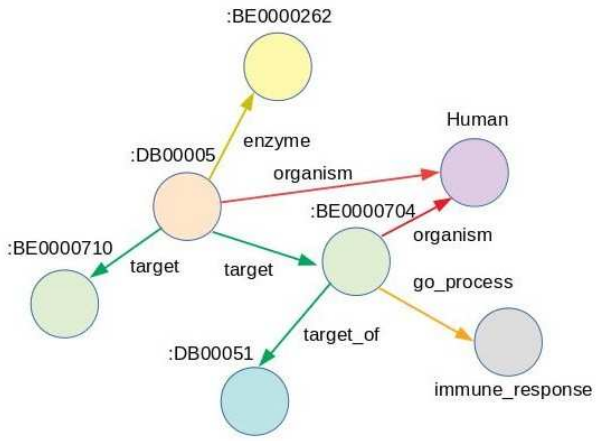

Fig. 2. A visualization of the structure of DrugBank Knowledge Graph.

IV. Methodology

Mathematically, a Knowledge Graph is defined as $G = (V, R, E)$ with entities $v_i \in V$, relations $r \in R$. A triple is *(head,relation,tail)* or $(h, r, t) \in E$ where $h \in V$, $t \in V$, $r \in R$. Link Prediction is the prediction of correctness of a new triple $(\hat{h}, r, \hat{t}) \notin E, \hat{h} \in V, \hat{t} \in V$. Link Prediction is thus the prediction of new facts which are not already known.

As described in section II, the basic relational learning model has first embedding the entities into d-dimensional vectors, for $(h, r, t) \in E$, this gives $\mathbf{h} \in \mathbb{R}^d, \mathbf{t} \in \mathbb{R}^d$ and the relational scoring function would then score the triple. The scoring function is different across different methods. The relational scoring function f_r gives score s for the (h, r, t) as $s(h, r, t) = f_r(\mathbf{h}, \mathbf{t})$, $s \in \mathbb{R}$.

In this work, we experimented with four different kind of scoring schemes. These are 1: Random Walk method (Neighborhood Preserving Method) 2: Translation based Scoring in Embedding Space 3: Bilinear Scoring Function 4: Convolutional Scoring Function. The latter three are proposed in [2], [6], and [8] respectively.

Random Walk Method

Random Walk method is inspired from *language-modeling* [18] in Natural Language Processing. In *language-model* inspired Random Walk method, the goal is to predict the next node v_i in a sequence of extracted some fixed length path(truncated Random Walk) $(v_1, ..., v_{i-2}, v_{i-1})$. In a vanilla recurrent neural network model, this is done as follows:

$$\phi_i = Ev_{i-1} \tag{1}$$

$$h_i = f(\phi_i, h_{i-1}) \tag{2}$$

$$y_i = softmax(Wh_i + b) \tag{3}$$

$$softmax = \frac{\mathrm{e^u}}{\sum_{\vee u \in V} \mathrm{e^u}} \tag{4}$$

v_i is the i_{th} vertex in a path generated, matrix $E \in \mathbb{R}^{d \times V}$ is the embedding matrix to be generated, *d* is the embedding dimension and *V* is the number of vertices in a graph. Function *f()* is LSTM [19] unit which takes ϕ_i, the representation of the current node and the previous hidden state and outputs the next hidden state. $Wh_i + b$ is basic Decoder unit which passes through softmax to output y_i, the probability distribution for the next node in the path. The loss is then calculated between the model prediction distribution y_i and the actual observed distribution for the next node in the path v_i^*(one-hot encoded distribution). We call this observed distribution as the target distribution. We used Cross-Entropy Loss to train the neural network.

$$L_i = CrossEntropy(y_i, v_i^*) \tag{5}$$

$$L_i = -\sum v_i^* \log y_i \tag{6}$$

This deep neural network learns complex relationships in the graph and also models non-linearities present. However, note that the concept of relations is lost in this model as every edge is considered of the same type while traversing the graph.

Translation based Scoring Function
This is a translational distance model. Here, the constrained maintained is $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, $\mathbf{r} \in \mathbb{R}^d$, i.e. *relation* vector when translated with *head* vector should approximately be equal to the *tail* vector.

$$s(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}|| \tag{7}$$

$||.||$ is l2-norm.

Bilinear Scoring Function
This is a semantic matching model. Bilinear scoring function is:

$$s(h, r, t) = \mathbf{h}^T f_r \mathbf{t}; f_r \in \mathbb{R}^{d \times d} \tag{8}$$

To reduce the number of parameters so as not to overfit, f_r is taken as diagonal matrix.

Convolutional Scoring Function
This is a deeper model which uses 2D Convolutional operator to model relationship between the *head* and *relation*. The scoring function

$$s(h, r, t) = g(vec(g([\bar{\mathbf{h}}; \bar{\mathbf{r}}] * w))W), \tag{9}$$

$\mathbf{r} \in \mathbb{R}^d$ is f_r, $\bar{\mathbf{h}}$ represents 2D reshaped $\mathbf{h}$, $*$ is convolutional operator and W is parameter matrix of fully-connected layer, g is non-linear function(ReLU).

V. Experiments and Results

A. Data Preparation

For evaluation purposes, the triples from the dataset are splitted into training,testing and validation data. Testing and Validation data have triples which are excluded from the training data to analyse the generalizing capacity of the models. However, note that for some relations, the triple count is less. So for all the drug entities, only those relations considered which have greater than five tail entities out of which two are randomly selected to be included in validation and testing data. This gives us 10,193 validation triples and 637 testing triples.

B. Evaluation metric for Link-Prediction

Conventionally, Link-Prediction is considered as an entity ranking task. Every entity is considered as a target entity for a triple in testing data. The scoring function will weigh true targets more than false targets for a triple. Evaluation metrics thus considered are Mean Average Precision(MAP) and Mean Reciprocal Rank(MRR).

TABLE II
Link Prediction Results

Model	MRR	MAP
Random Walk Method	0.12	0.27
Translation Based Model	0.29	0.45
Bilinear Scoring Model	0.43	0.59
Convolutional Scoring Model	0.24	0.38

C. Experimental-Setup

For Random Walk method, LSTM has 3 hidden layers with 200 hidden units per layer. The embedding dimension is set to 100 and all the initial parameters are initialized using uniform distribution. Model is trained using Adam Optimizer [20] with 0.001 initial learning-rate for 100 epochs at a batch-size of 256. Model strength is increased until the good performance is obtained at 3 hidden layers and 200 hidden units. The walks length for truncated random walk is 80 and for every node we consider 10 walks. The next node is decided by uniform probability distribution and the model is trained using truncated Backpropagation at every 10 steps to avoid vanishing gradients.

For Translation based model, embedding dimension is 100 and L2 norm is considered. Margin-Ranking Loss is minimized with margin at1.0, and the learning rate is 0.01, model is trained with SGD with momentum optimizer, momentum = 0.9.

For Bilinear scoring model, batch size = 32, for every true triple 10 corrupted triples are sampled, embedding dimension = 100, Adagrad Optimizer [21] for training is used with weight decay of 0.00001 for regularization and is trained for 100 epochs.

For Convolutional scoring model, hyperparameters are same as the oiginal paper with batch size being 216 and training for 100 epochs. The models are trained with performance monitoring on validation dataset.

D. Results

The results for Link Prediction are shown in II. As can be seen, Bilinear scoring model has superior performance compared to other models. Since the concept of relation criteria is not present in Random Walk method, the poor performance is explainable. But despite being deeper model, Convolutional scoring model is poor as compared to shallow models like Bilinear scoring model and Translation Based Model. This can be attributed to the fact that the average degree of a node in the graph considered is 6.41 and as per [6] a shallow model is sufficient is accurately represent the structure of the graph.

E. Efficacy of the Representation Learning to Drug Informatics

We further consider two tasks to check the strength of the representation learning to study Drugs. These are 1) Drug-Drug Interaction Prediction(to predict if the combination of two drugs is going to interact) 2) Drug-Target Prediction. The learned representation of the drugs, proteins, etc. are the features and the advantage of this method is that we are having an automated feature learning.

Both tasks are supervised learning binary-classification tasks where for a drug-drug pair and drug-target pair, we have to find whether that pair is True of False. To generate the negative(false) examples: for every true pair, a false pair is randomly sampled. The classifier used is Support Vector Machines as implemented in libsvm [22] using default parameters. For a pair, the feature vectors of its individual components are concatenated together. Metric used is F1-score and the results are shown in III. Note that since Bilinear Scoring Model has superior performance over other models, we used these Representations only. To compare these results with other systems in literature is difficult as the experimental settings are different. However, [23] employed Knowledge Graph for structural and textual similarity to predict drug-drug interactions and the published F1 score is 0.85 with all features included. This representation learning based approach is therefore, more-or-less comparable to their work.

TABLE III
DRUG-DRUG INTERACTION AND DRUG-TARGET PREDICTION RESULTS(FROM DISTMULT MODEL)

Task	F1-score
Drug-Drug Interaction Prediction	0.87
Drug-Target Prediction	0.81

VI. CONCLUSION

In this work, we have used representation learning on knowledge graph on the openly available large-scale DrugBank Knowledge Graph. We experimented with four different methods. This method gives us dense representations for entities like drugs, targets, etc. which can be used to perform further machine learning tasks. We also show how the feature representations learned as part of the process can be used to perform regular tasks in machine learning based drug study. We perform two tasks; drug-drug interaction prediction and drug-target prediction and got good performance.

REFERENCES

[1] Quan Wang, Zhendong Mao, Bin Wang, Li Guo, "Knowledge Graph Embedding: A Survey of Approaches and Applications",IEEE Transactions on Knowledge and Data Engineering vol.29, no. 12, 2017, pp. 2724-2743.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, Oksana Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data" Neural Information Processing Systems (NIPS),2013, South Lake Tahoe, United States, pp.1-9.

[3] Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," 29th AAAI Conference on Artificial Intelligence,2015,pp. 2181-2187.

[4] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embed-ding via dynamic mapping matrix," 53rd Annual Meeting Associations of Computational Linguistics 7th International Joint Conference on Natural Language Processing,2015, pp. 687-696.

[5] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model forcollective learning on multi-relational data", 28th International Conference on Machine Learning, 2011, pp. 809-816.

[6] B. Yang, W.-T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases", International Conference on Learning Representations, 2015.

[7] R. Socher, D. Chen, C. D. Manning, and A. Y. Ng, "Reasoning with neural tensor networks for knowledge base completion", Adv. Neural Inf. Process. Syst.,2013, pp. 926-934.

[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel,"Convolutional 2D Knowledge Graph Embeddings", Proceedings of the 32th AAAI Conference on Artificial Intelligence, 2018, pp. 1811-1818.

[9] Ristoski, Petar, and Heiko Paulheim. "Rdf2vec: Rdf graph embeddings for data mining", International Semantic Web Conference. Springer, Cham, 2016, pp. 498-514.

[10] Bryan Perozzi, Rami Al-Rfou, Steven Skiena,"DeepWalk: Online Learning of Social Representations",Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York USA, 2014,pp. 701-710.

[11] Aditya Grover,Jure Leskovec,"Node2vec: Scalable Feature Learning for Networks",Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[12] Vilar, Santiago, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, Nicholas P. Tatonetti, "Similarity-based modeling in large-scale prediction of drug-drug interactions", Nature protocols 9, no. 9, 2014, pp. 2147.

[13] Zhang, Wen, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, Xiaohong Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data", BMC bioinformatics 18, no. 1, 2017, pp. 18.

[14] Cheng, Feixiong, Zhongming Zhao, "Machine learning-based prediction of drugdrug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties", Journal of the American Medical Informatics Association 21, no. e2, 2014, pp. e278-e286.

[15] Gottlieb, Assaf, Gideon Y. Stein, Yoram Oron, Eytan Ruppin, Roded Sharan, "INDI: a computational framework for inferring drug interactions and their associated recommendations", Molecular systems biology 8, no. 1, 2012, p. 592.

[16] Wishart, David S., Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, Murtaza Hassanali, "DrugBank: a knowledge base for drugs, drug actions and drug targets", Nucleic acids research 36, 2007, pp. D901-D906.

[17] Callahan, Alison, Jos Cruz-Toledo, Peter Ansell, Michel Dumontier, "Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data", In Extended Semantic Web Conference Springer, Berlin, Heidelberg, 2013, pp. 200-212.

[18] Yoshua Bengio, "A neural probabilistic language model", Journal of machine learning research 3,2013, pp. 1137-1155.

[19] Hochreiter, Sepp, Jrgen Schmidhuber, "Long short-term memory", Neural computation 9.8 1997, pp. 1735-1780.

[20] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization", arXiv preprint, 2014, arXiv:1412.6980

[21] Duchi, John, Elad Hazan,Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization", Journal of Machine Learning Research 12.Jul, 2011, pp. 2121-2159.

[22] Chang, Chih-Chung, Chih-Jen Lin, "LIBSVM: A library for support vector machines",ACM transactions on intelligent systems and technology (TIST) 2, no.3, 2011, pp. 27.

[23] Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, Mohammad Sadoghi,"Large-scale structural and textual similarity-based mining of knowledge graph to predict drug- drug interactions", Web Semantics: Science, Services and Agents on the World Wide Web 44, 2017, pp. 104-117.