Knowledge Graph Representation Learning based Drug Informatics

Rajeev Verma and Dr. Preetam Kumar

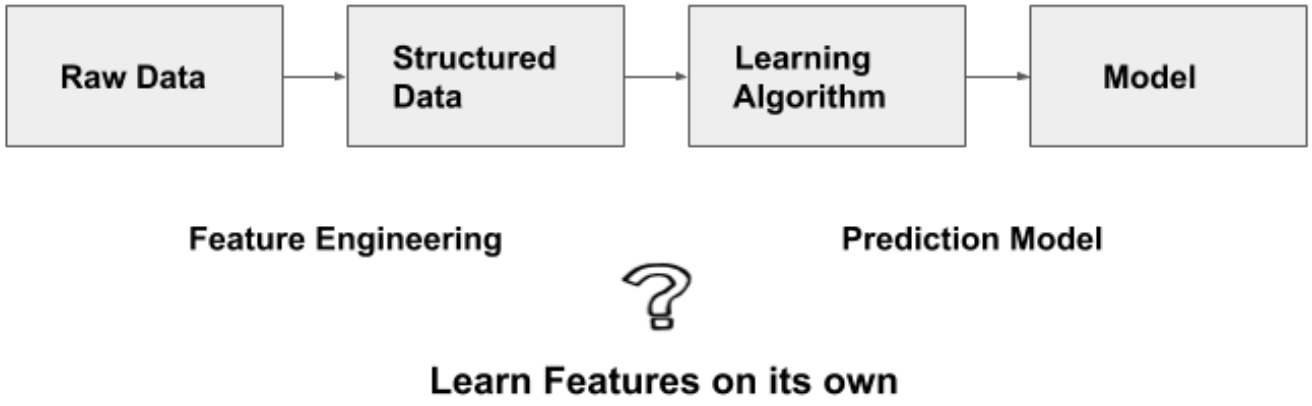
IIT Patna

July 26, 2019

Outline

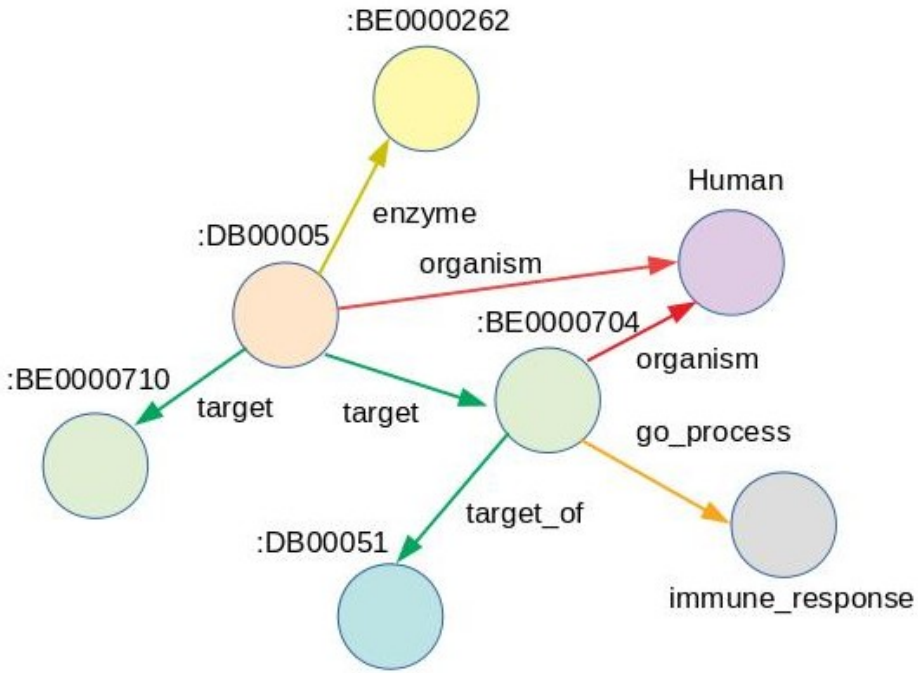
Outline

Motivation



- The performance of machine learning system is affected by the data representation technique (Feature Engineering).
- Automated Feature Learning(Representation Learning) is the major strength of the modern Deep Learning systems.
- Knowledge Graphs are large-scale knowledge bases containing real world facts which can be harnessed to perform data-intensive studies.

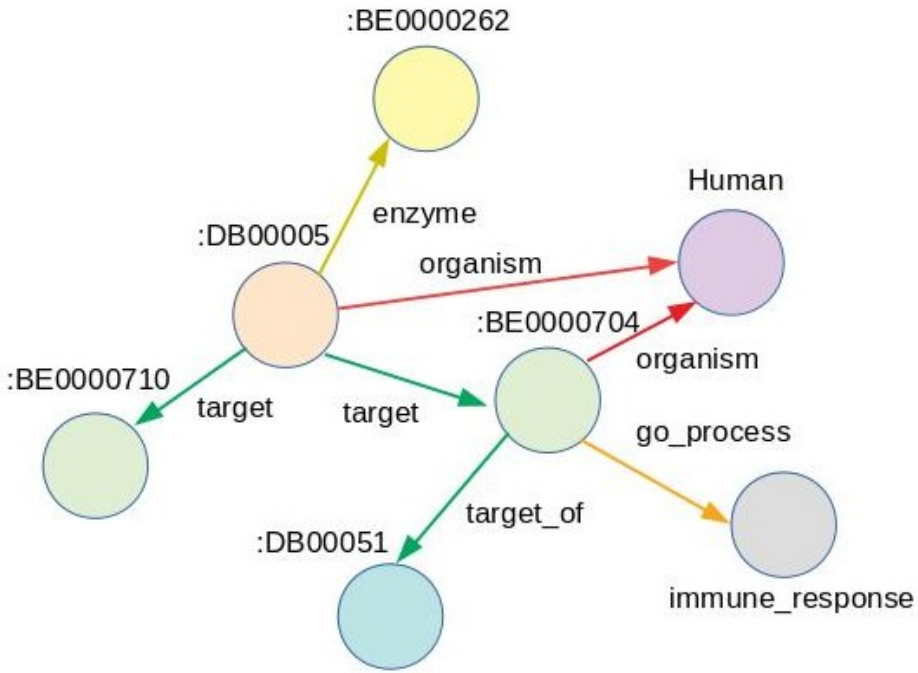
Knowledge Graph - Definition



A visualization of the structure of DrugBank Knowledge Graph. Example of the fact included: *(:BE0000262, enzyme, :DB00005)*, representing that tail is the enzyme of the *head*.

- Multi-Relational data where knowledge is represented as triple *(head,relation,tail)*, *head* and *tail* are entities and *relation* specifies how two are related through the directed edge.

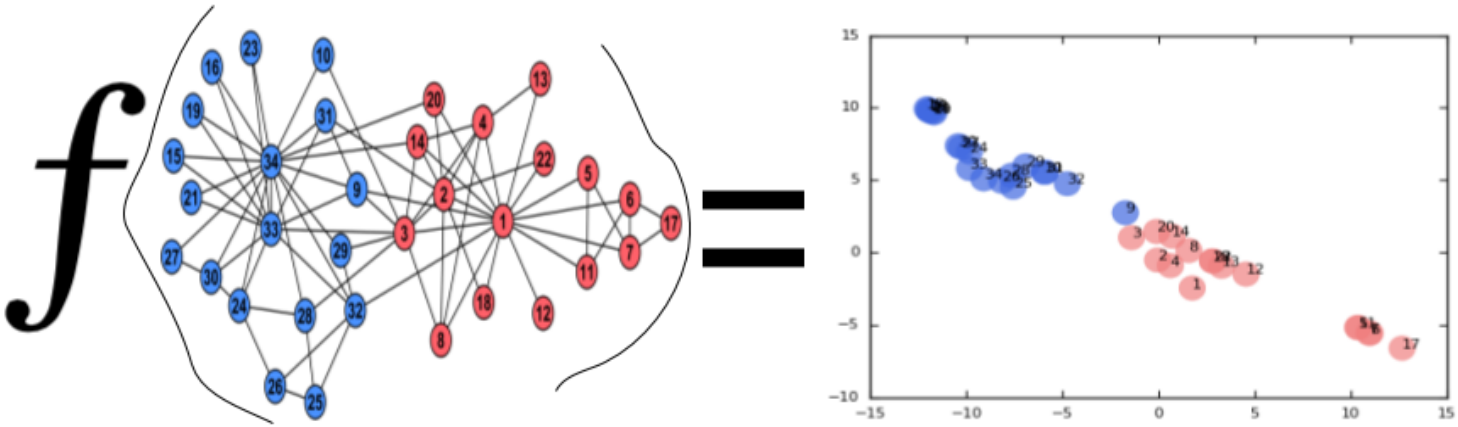
Knowledge Graph - Definition



A visualization of the structure of DrugBank Knowledge Graph. Example of the fact included: *(:BE0000262, enzyme, :DB00005)*, representing that tail is the enzyme of the *head*.

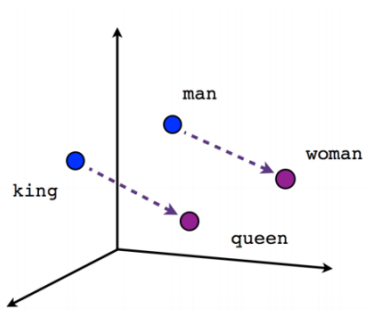
- Multi-Relational data where knowledge is represented as triple *(head,relation,tail)*, *head* and *tail* are entities and *relation* specifies how two are related through the directed edge.
- Mathematically, $G = (V, R, E)$ with entities $v_i \in V$, relations $r \in R$ and triple *(head,relation,tail)* or $(h,r,t) \in E$ with $h \in V, r \in R$ and $t \in V$.

Representation Learning in Graphs

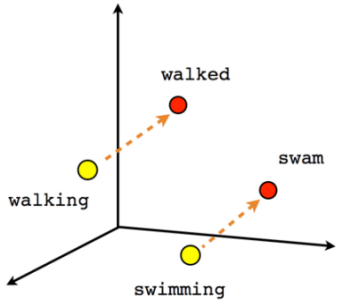


- Representation Learning in Knowledge Graph aims to embed the components of the knowledge graph(entities and relations) to some $d-$dimensional continuous vector space, preserving the structure and semantics of the Knowledge graph.

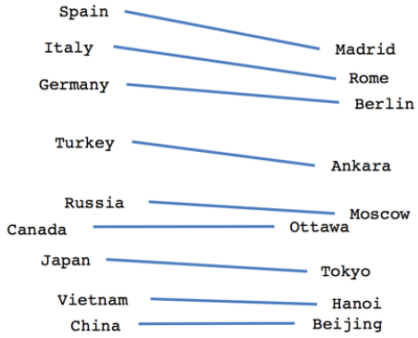
Representation Learning in Graphs



Male-Female



Verb tense



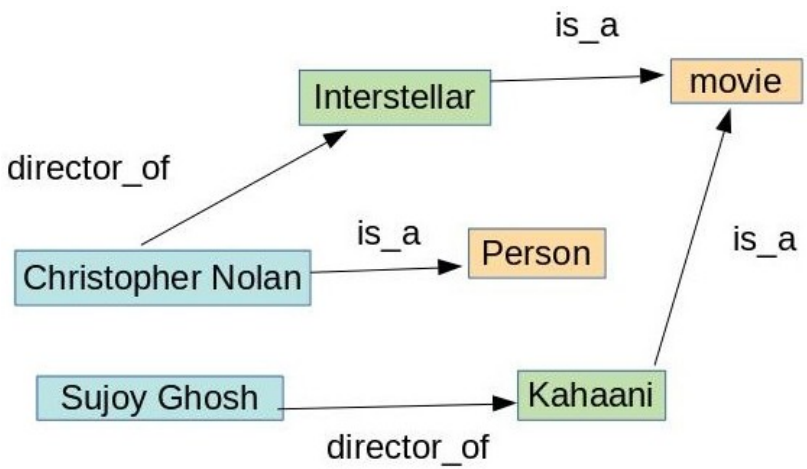
Country-Capital

- These d-dimensional vectors (also known as embeddings) capture semantics of the real-world concept which can be used for performing downstream tasks.
- Representation Learning on words revolutionized the entire paradigm of Natural Language Processing systems and made them ubiquitous as they are today.[1] [2]

[1] https://www.tensorflow.org/tutorials/representation/word2vec

[2] Distributed Representations of Words and Phrases and their Compositionality, Mikolov et al.

Why this is even more important?



- Data Sparsity and Data Incompleteness affect the performance of systems using these Knowledge Bases for downstream tasks.
- 66% People in DBpedia Knowledge Graph do not have Date of Birth information, 58% of the scientists have missing field of working.[3]
- Statistical Relational Learning: Automatic Discovery of new facts, thus completing the knowledge base.
- The goal in representation Learning is to embed in such vector spaces where the embedded entity can reason meaningfully about the missing information.
- The graph-structure of the Knowledge Graph contains information about the entity through neighborhood and locality structure.

[3]Type-Constrained Representation Learning in Knowledge Graphs, Denis KrompaB et al.

DrugBank and Drug Informatics

- Drug Discovery and assessment - a complex but an acute process.
- Current approaches to this end involve clinical evaluation and post-marketing surveillance - expensive and time-consuming.

DrugBank and Drug Informatics

- Drug Discovery and assessment - a complex but an acute process.
- Current approaches to this end involve clinical evaluation and post-marketing surveillance - expensive and time-consuming.
- The DrugBank database is a comprehensive, freely accessible, online database containing information on drugs and drug targets. As both a bioinformatics and a cheminformatics resource, DrugBank combines detailed drug data with comprehensive drug target information. Maintained by Canadian Institutes of Health Research.
- This representation Learning based data-intensive approach is motivated to make use of this available knowledge to accelerate the process of drug-discovery.
- By meaningful embeddings of the entities in the knowledge base, we can use those embeddings as learned features to perform machine learning tasks.

Outline

General Approach to Learning Representations

Given a graph $G = (V, R, E)$, there are two main components: Encoder module and Scoring function.

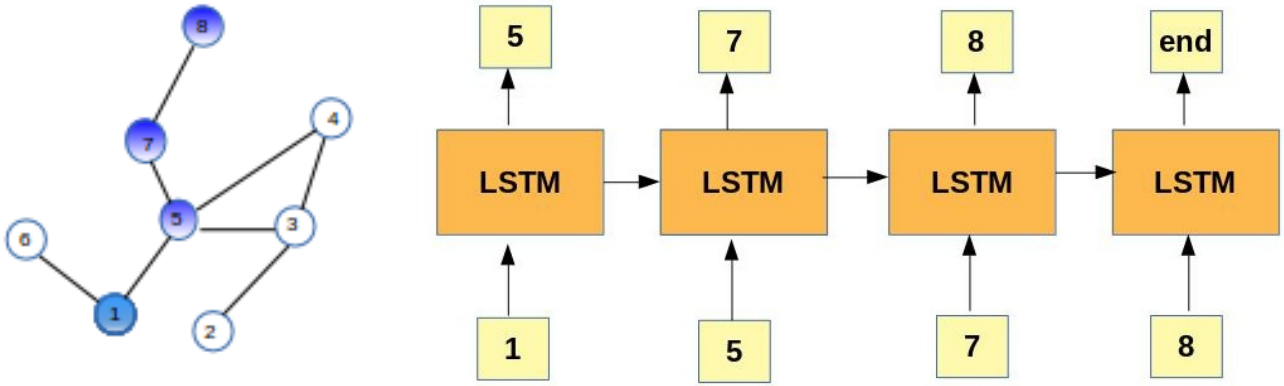
- **Encoder Module:** For a triple $(h, r, t) \in E$, this gives $\mathbf{h} \in \mathbb{R}^d, \mathbf{t} \in \mathbb{R}^d$.
- **Scoring Function:** Function f_r scores the triples as $s((h, r, t)) = f_r(h, t), s \in \mathbb{R}$.
- f_r is different according to different scoring schemes.

In this work, we experimented with four different kind of scoring schemes:

- Similarity Based Scoring
- Translation Based Scoring
- Bilinear Scoring Function
- Convolutional Scoring Function

Random Walk Method (Similarity Based Scoring)

An entity is known by the company it keeps.



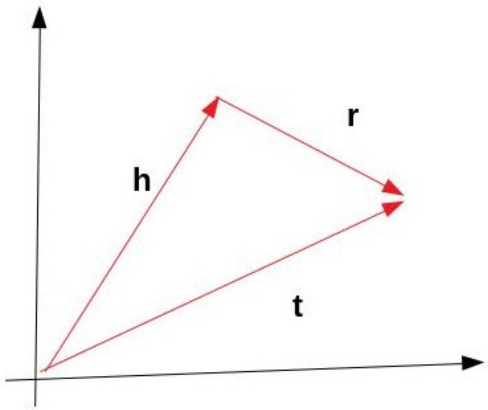
A sampling of the truncated random walk from the graph, $(1 \Rightarrow 5 \Rightarrow 7 \Rightarrow 8)$ and the random walk method in action. Given the current node, the model predicts the next node in the walk. Note that the circle represents the node while the square is the $d-$dimensional vector of the node.

- Given a node $v \in V$, learn embedding $\mathbf{v} \in \mathbb{R}^d$ which is predictive of the nodes in neighborhood of v, $N_R(v)$, i.e.

$$max \sum_{v \in V} logP(N_R(v)|\mathbf{v})$$

Dataset: Starting with every node, simulate truncated random walks in the graph;
for *every random walk* $(v_1, v_2, ..., v_{i-1})$ **do**
  Predict the v_i node as;
  $\mathbf{v}_{i-1} = \mathbf{E}v_{i-1}$;
  $\mathbf{h}_i = f(\mathbf{v}_{i-1}, \mathbf{h}_{i-1})$;
  $\mathbf{y}_i = softmax(W\mathbf{h}_i + b)$;
  Calculate the loss between $\mathbf{y}_i$ and the actual target distribution $\mathbf{v}_i^*$;
  $L_i = -\sum \mathbf{v}_i^* log(\mathbf{y}_i)$;
  Optimize;
end

Translation Based Scoring



- Poses a distance based constraint in the embedding space.
- for a triple (h, r, t), it maitains that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, $\mathbf{h} \in \mathbb{R}^d, \mathbf{r} \in \mathbb{R}^d, \mathbf{t} \in \mathbb{R}^d$.
- The scoring function is $s((h, , r, t)) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$.
- This scoring gives score to the true triples in the knowledge graph.
- The set $\hat{E}$ of false triples is obtained as:

$$\hat{E} = \Big\{(\hat{h}, r, t)|(\hat{h} \in V, (\hat{h}, r, t) \notin E)\Big\} \cup \Big\{(h, r, \hat{t})|(\hat{t} \in V, (h, r, \hat{t}) \notin E)$$

- Training is done using Margin Ranking Loss as:

$$Loss = \sum_{(h,r,t) \in E} \sum_{(\hat{h},r,\hat{t}) \in \hat{E}} max\Big(0, 1 + s(h, r, t) - s(\hat{h}, r, \hat{t})\Big)$$

Translating Embeddings for Modeling Multi-relational Data, Bordes et al.

Bilinear Scoring Function

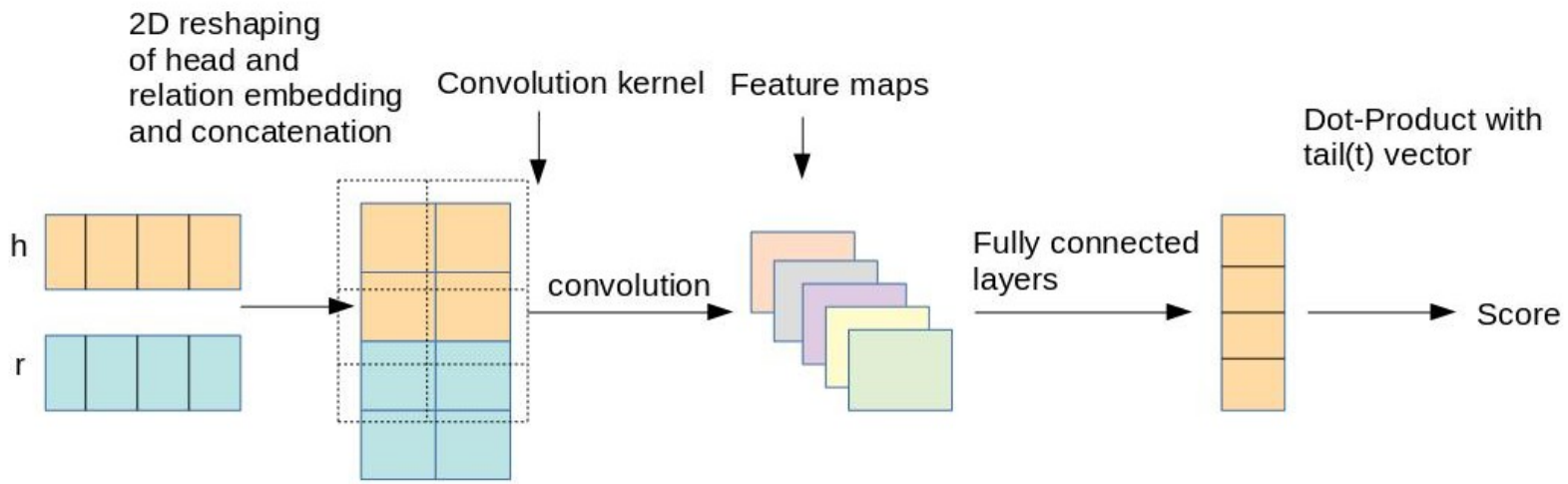
- For a triple (h, r, t), a bilinear transformation is used as the scoring function.

$$s((h, r, t)) = \mathbf{h}^T f_r \mathbf{t}$$

$$f_r \in \mathbb{R}^{d \times d}, \mathbf{h} \in \mathbb{R}^d, \mathbf{t} \in \mathbb{R}^d$$

- Every relation $r \in R$ has been embedded as $(d \times d)$ matrix.
- True Triples should score high according to the bilinear transformation.

Convolutional Scoring Function



- To score the triple *(h,r,t)*, the representations $\mathbf{h} \in \mathbb{R}^d$, $\mathbf{r} \in \mathbb{R}^d$, and $\mathbf{t} \in \mathbb{R}^d$ are first randomly initialized.
- The $\mathbf{h}$ and $\mathbf{r}$ are then reshaped to give $\bar{\mathbf{h}}$ and $\bar{\mathbf{r}}$, $\bar{\mathbf{h}} \in \mathbb{R}^{d_w \times d_h}$, $\bar{\mathbf{r}} \in \mathbb{R}^{d_w \times d_h}$, which are further concatenated to form a image sort of 2D representation. Here, $d = d_w d_h$.
- On this representation, regular convolutional filters are applied to give feature maps which are then subsampled and by passing through fully connected layers, gives a vector $\tilde{\mathbf{t}} \in \mathbb{R}^d$.
- We take dot-product of the $\tilde{\mathbf{t}}$ and the $\mathbf{t}$ vector which gives score of the triple, $s(h, r, t)$. Mathematically, the entire operation is given as

$$s(h, r, t) = f(f(c * k)W)\mathbf{t}$$

$c = [\tilde{\mathbf{h}}; \tilde{\mathbf{r}}]$, $*$ is the convolution operator, k is the kernel.

Outline

Evaluation Metric

The evaluation of the learned representations is done using Link Prediction task, where the training is done using only a sample of the true triples and the correctness of the unseen triples is predicted. This evaluation is inspired from document-retrieval task in Information Retrieval.

Mean Average Precision

- For the query (h, r), all the $t_s \in V$ are scored.
- True triples should be at the top while the false ones should be at the bottom. Accordingly, Precision is defined as

$$Precision = \frac{|\{trueentities\} \cap \{entitiesretrieved\}|}{|\{entitiesretrieved\}|}$$

Average Precision: when Precision is calculated for every position in the retrieved list. Mean Average Precision: average of average precision over all queries.

Mean Reciprocal Rank

- The true triples in the retrieved list should have top ranks as compared to the false ones. Mean Reciprocal rank is defined as

$$MeanReciprocalRank = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{Rank(n)}$$

N is the total number of queries.

Knowledge Graph Description

We used DrugBank Data from the Linked Open Data, Bio2RDF. Bio2RDF: open-source project, creates large RDF graph interli nking data from biological datasets.

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

Table 1: Dataset Statistics

The common relations are *drug-classification category*, *calculated-properties*, *ddi-interactor-in*, *ingredient*, *target*, *enzyme*, *mechanism-of-action*, *protein-binding*, *toxicity*, *gene-sequence*, etc.

Note that the average degree of a node in the graph is 6.41.

Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data, Callahan et al.

Results on Link Prediction

- The triples from the dataset are splitted into training,testing and validation data.
- 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.

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

Table 2: Link Prediction Results

Outline

Tasks

- We check the efficacy of the learned representations to two drug study tasks: **Drug-Drug Interaction Prediction** and **Drug-Target Prediction**.
- The learned representations are used as features (Automated feature learning).
- Tasks are supervised learning binary classification task.
- For every true example, a false pair is randomly sampled.
- Trained using Support Vector Machines.
- For a pair, the feature vectors of its components are concatenated together.

Results

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

Table 3: Drug-Drug Interaction and Drug-Target Prediction Results(from Bilinear model)

- To compare these results with other systems in literature is difficult as the experimental settings are different.
- However, the reported work on somewhat similar settings which employed Knowledge Graph for structural and textual similarity(along with embeddings) to predict drug-drug interactions and the published F1 score is 0.85 with all features included.
- We can say that this work is promisingly comparable.

Large-scale structural and textual similarity-based mining of knowledge graph to predict drug- drug interactions, Abdelaziz et al.

Outline

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.

Research on Representation Learning in Network Biology

- Representation Learning in Networks and Knowledge Graphs is currently one of the most researched fields in Deep Network Learning.
- The research is focused on building more meaningful and efficient representations system(like Graph Convolutional Networks[7]).
- These systems have shown powerful results in Social Networks like Recommender System, Community Detection, Question Answering, etc.
- This research can be extended to Network Biology and Bioinformatics to accelerate the biological science.[8] [9]

[7] Semi-Supervised Classification with Graph Convolutional Networks, Thomas N. Kipf, Max Welling

[8] To Embed or Not: Network Embedding as a Paradigm in Computational Biology, W. Nelson, M. Zitnik, B. Wang, J. Leskovec, A. Goldenberg, R. Sharan

[9] Modeling Polypharmacy Side Effects with Graph Convolutional Networks. M. Zitnik, M. Agrawal, J. Leskovec