Predicting Missing Links Using PyKEEN

Mehdi Ali[1,2], Charles Tapley Hoyt[3], Daniel Domingo-Fernández[3], and Jens Lehmann[1,2]

Smart Data Analytics Group, University of Bonn, Germany
{mehdi.ali,jens.lehmann}@cs.uni-bonn.de
Department of Enterprise Information Systems, Fraunhofer Institute for Intelligent Analysis and Information Systems, Sankt Augustin and Dresden, Germany
jens.lehmann@iais.fraunhofer.de
Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany
{charles.hoyt,daniel.domingo.fernandez}@scai.fraunhofer.de

Abstract. PyKEEN is a framework, which integrates several approaches to compute knowledge graph embeddings (KGEs). We demonstrate the usage of PyKEEN in an biomedical use case, i.e. we trained and evaluated several KGE models on a biological knowledge graph containing genes' annotations to pathways and pathway hierarchies from well-known databases. We used the best performing model to predict new links and present an evaluation in collaboration with a domain expert*.

Keywords: Machine learning · Knowledge Graphs · Bioinformatics · Link Prediction

1 Introduction

Knowledge graphs (KGs) have been adopted by various research fields (e.g., Semantic Web, bioinformatics) to represent factual information. Examples of KGs are DBpedia [7], Wikidata [13], and the Bio2RDF [2] repository. Although existing KGs may contain billions of links, they are usually incomplete (i.e., missing links) [9]. Knowledge graph embeddings (KGEs), which learn latent vector representations for entities and relations in KGs while best preserving their structural characteristics, provide one avenue for predicting these missing links.

Because the software ecosystem for KGEs remains limited, we have developed the KEEN Universe [1] for training, evaluating, and sharing KGEs with a strong focus on reproducibility and transferability. It currently comprises the Python packages: PyKEEN (Python KnowlEdge EmbeddiNgs), BioKEEN (Biological KnowlEdge EmbeddiNgs), and the KEEN Model Zoo for sharing experimental artifacts.

In this demonstration paper, we present a link prediction use case from the biomedical domain that accompanies our resource paper at the ISWC 2019 Conference [1]. In particular, we focus on the use of PyKEEN in predicting missing links between genes and biological pathways as well as their internal hierarchies.

2 PyKEEN

PyKEEN provides the functionalities to train and to evaluate KGEs, and it provides an inference workflow that assists users to predict novel links. PyKEEN consists of two layers: the *configuration layer* and the *learning layer*.

Configuration Layer The configuration layer assists users in specifying their KGE experiments' datasets, models, hyper-parameters, training procedures, and evaluation procedures. Experiments can either be defined programmatically or by using the interactive command line interface (CLI) via a terminal. The CLI ensures that experiments are configured correctly, and in case users provide an invalid input, the CLI informs the users and provides an example of a correct input.

Learning Layer The *learning layer* trains a model in *training mode* based on a user defined set of hyper-parameters or finds suitable hyper-parameter values in *hyper-parameter optimization (HPO) mode*.

Inference Workflow The inference workflow generates for a user defined set of entities and relations all possible triple permutations (users can specify to exclude reflexive triples of the form *(e,r,e)*). The inference workflow exports a file containing the triples and their predicted scores where the predictions are sorted according to their scores.

3 Application

Biological pathway databases have been generated and used in the classical analysis of -*omics* data, but their formulations as KGs are not amenable to classical machine learning approaches for classification, clustering, or predictive modeling. Here, we trained and evaluated three KGE models (i.e., TransE, TransR and ComplEx) before selecting the best performer for prediction of novel roles of genes in pathways and evaluation by a domain expert.

Training was conducted using four datasets: three that comprise links between genes and pathways from disparate pathway databases (i.e., KEGG [6], Reactome [5], and WikiPathways [11]) and one (i.e., ComPath [4]) that comprises manually curated links between pathways from the previously mentioned resources. By predicting links in the resulting merged KG, we identified and hypothesized the role of genes in novel pathways.

Experimental Setup For each experiment, we split the KG into a training and test set then performed HPO for the TransE [3], TransR [8], and ComplEx [12] models. The results were evaluated by *mean rank* and *hits@k* and presented in Table 1. Afterwards, we focused on a set of entities and the relation *partOf* and considered triples of the form *(gene, partOf, pathway)* to predict novel links using PyKEEN's inference workflow that were evaluated by a domain expert.

Model	Mean Rank	Hits@10
TransE [3]	1069.21	13.88%
TransR [8]	376.38	24.83%
ComplEx [12]	193.98	59.50%

Table 1. HPO results.

While ComplEx performed well, the poor performance of TransE may be due its poor abilities to handle the high cardinality (N-M) relations in the data. Due to time constraints, only 5 iterations of HPO were performed for TransR, so its results may also be improved.

Results Due to the large number of predictions made by the KGE model, we focus on the top five predictions between genes and pathways in Table 2. By looking at these highly plausible links, we can not only identify novel roles of genes in pathways, but also hypothesize the role of pathways in diseases that has been linked to a given gene.

Gene	Database	Pathway	Score
RXRA	WikiPathways	Nuclear Receptors in Lipid Metabolism and Toxicity	16.20
UPP1	KEGG	Pyrimidine metabolism	16.01
EZR	KEGG	Shigellosis	13.63
UGT1A1	KEGG	Porphyrin and chlorophyll metabolism	13.41
BLM	WikiPathways	DNA IR-damage and cellular response via ATR	12.53

Table 2. Top ten predicted gene-pathway links in which higher scores indicate more plausible links.

The two most confidence predictions suggest that *RXRA* and *UPP1* play a role in *Nuclear Receptors in Lipid Metabolism and Toxicity* and *Pyrimidine metabolism* pathways, respectively. A survey of the recent biomedical literature suggests that *RXRA* is involved in lipid metabolism and *UPP1* in the degradation and salvage of pyrimidine ribonucleosides. Interestingly, the third predicted link that suggests the involvement of *EZR*, a cytoplasmic peripheral gene, in the disease Shigellosis has been previously described by [10] in which they implicated the gene in the process of Shigella bacterial uptake. Ultimately, it could

be interesting to investigate other possible links connecting genes to pathway implicated in diseases.

4 Conclusion

We have demonstrated the usage of PyKEEN in predicting missing links in KGs from the biomedical domain. In particular, we performed HPO for three KGE models (i.e., TransE, TransR, and ComplEx) and selected the best performing to provide predictions to a domain expert which manually evaluated the top ranked predictions. Finally, that this workflow that can be applied in any domain highlights the effectiveness of PyKEEN to discover novel knowledge.

Acknowledgement This work was supported by the German national funded BmBF project MLwin.

References

1. Ali, M., et al.: The keen universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability. In: International Semantic Web Conference (2019)
2. Belleau, F., et al.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics **41**(5), 706–716 (2008)
3. Bordes, A., et al.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
4. Domingo-Fernandez, D., et al.: ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. npj Systems Biology and Applications **4**(44) (2018). https://doi.org/10.1038/s41540-018-0078-8
5. Fabregat, A., et al.: The Reactome Pathway Knowledgebase. Nucleic Acids Research **46**(D1), D649–D655 (jan 2018). https://doi.org/10.1093/nar/gkx1132
6. Kanehisa, M., et al.: Kegg: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research **45**(D1), D353–D361 (2017)
7. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal **6**(2), 167–195 (2015), outstanding Paper Award (Best 2014 SWJ Paper)
8. Lin, Y., et al.: Learning entity and relation embeddings for knowledge graph completion. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
9. Nickel, M., Murphy, et al.: A review of relational machine learning for knowledge graphs. Proceedings of the IEEE **104**(1), 11–33 (2015)
10. Skoudy, A., et al.: A functional role for ezrin during shigella flexneri entry into epithelial cells. Journal of Cell Science **112**(13), 2059–2068 (1999), https://jcs.biologists.org/content/112/13/2059
11. Slenter, D., et al.: WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic acids research **46**(D1), D661–D667 (jan 2018). https://doi.org/10.1093/nar/gkx1064
12. Trouillon, T., et al.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning. pp. 2071–2080 (2016)
13. Vrandečić, D., et al.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)