Bioinformatics, 36(13), 2020, 4097–4098
doi: 10.1093/bioinformatics/btaa274
Advance Access Publication Date: 27 April 2020
Applications Note

Data and text mining

OpenBioLink: a benchmarking framework for large-scale biomedical link prediction

Anna Breit, Simon Ott, Asan Agibetov and Matthias Samwald *

Section for Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna 1090, Austria

*To whom correspondence should be addressed.

Associate Editor: Zhiyong Lu

Received on November 28, 2019; revised on April 3, 2020; editorial decision on April 20, 2020; accepted on April 21, 2020

Abstract

Summary: Recently, novel machine-learning algorithms have shown potential for predicting undiscovered links in biomedical knowledge networks. However, dedicated benchmarks for measuring algorithmic progress have not yet emerged. With OpenBioLink, we introduce a large-scale, high-quality and highly challenging biomedical link prediction benchmark to transparently and reproducibly evaluate such algorithms. Furthermore, we present preliminary baseline evaluation results.

Availability and implementation: Source code and data are openly available at https://github.com/OpenBioLink/OpenBioLink.

Contact: matthias.samwald@meduniwien.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in deep learning and vector-space embedding models have enabled the creation of a sizeable array of novel methodologies for link prediction—the task of predicting missing links in knowledge graphs. As many fundamental biomedical problems can be formulated as link prediction problems, there is growing interest in the application of these algorithms in the domain of biomedicine.

Advances in methodology are both measured and steered by established general-domain benchmarks, such as the FB15K benchmark derived from Freebase, the WN18 benchmark derived from WordNet (Bordes *et al.*, 2013) or the Unified Medical Language System (UMLS) benchmark (Dettmers *et al.*, 2018).

Unfortunately, these benchmarks are often found to have flaws such as information leakage between train and test sets (Toutanova and Chen, 2015) and do not reflect the domain-specific properties of heterogeneous biomedical knowledge bases. Instead of capturing primarily knowledge networks (FB15K) or hierarchical taxonomies (WN18, UMLS), biomedical knowledge bases often combine richly structured ontological hierarchies with large interaction networks. Predictions of interest usually cannot be made based on simple, crisp rules. Finally, biomedical knowledge graphs tend to be large, calling into doubt whether results from smaller benchmarks are informative.

It is not straightforward to adopt existing biomedical knowledge graphs such as Bio2RDF (Dumontier *et al.*, 2014) as benchmark datasets, as they contain a significant number of metadata relations that can interfere with the performance of link prediction algorithms, and special care has to be taken to exclude trivially inferable statements from the test set.

The first major work on embedding-based link prediction in the biomedical domain was published by Alshahrani *et al.* (2017). They evaluated a modified version of the DeepWalk algorithm adapted to heterogeneous graphs on a large-scale biomedical graph, containing Linked Data, biomedical ontologies and ontology-based annotations. Unfortunately, no benchmark dataset was established.

Crichton *et al.* (2017) and Yue *et al.* (2020) performed multiple evaluations on different graph embedding methods for link prediction, including different datasets and different train–test set splitting techniques. This work did not focus on evaluation of heterogeneous, multirelational graph data and corresponding algorithms. Recently PyKEEN (Ali *et al.*, 2019), a Python library for training and evaluation of link prediction methods, was introduced. It offers an excellent unified interface for various graph embedding models, but no dedicated benchmark dataset was established.

A dedicated, high-quality and highly challenging benchmark optimized for the task of evaluating link prediction methods in large, heterogeneous biomedical knowledge bases has not yet been established. In this article, we introduce the OpenBioLink suite of software, datasets and benchmarks to close this gap and to provide a highly transparent, reproducible and configurable evaluation framework.

2 Software architecture

The OpenBioLink framework consists of three modules: (i) graph creation module, (ii) train–test-split creation module and (iii) training and evaluation module. The user can interact with these modules

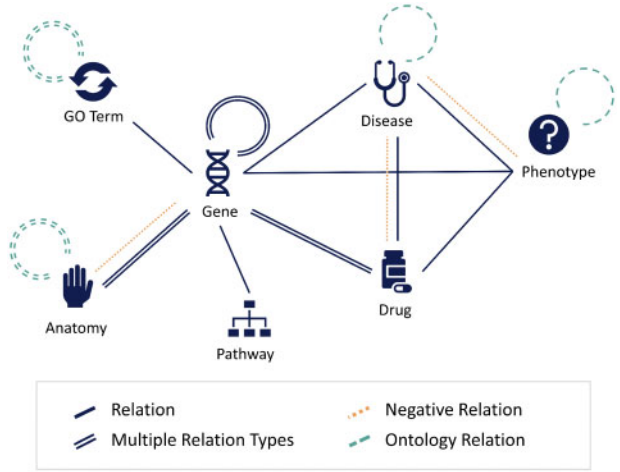

Fig. 1. An overview of the OpenBioLink benchmark graph

individually or use them together in a pipeline. The graph creation module creates the benchmark dataset from multiple public data sources. It allows for the creation of subsets and variations of the original benchmark, including the adaptation of the directionality and quality cut-off of edges as well as the exclusion of source databases or edge types. The train–test-split module divides data into a training and a test set, either randomly or via time slices. Special focus was put on the robustness and difficulty of the test set, which contains only entities that are also present in the training set and does not contain relations that can be trivially inferred from the training set (e.g. reverse edges of symmetric relations, inverse relations or super-relations). Negative samples are produced using the negative edges present in the benchmark data sources and—where needed—by applying typed negative sampling. In the third module, a model can be trained and tested. Models can be trained with external graph embedding libraries. Currently, an interface for PyKEEN is available. For evaluation, a wide range of metrics is offered, such as hits@k, mean reciprocal rank (MRR), area under the receiver operator characteristic curve (ROC AUC) and area under the precision-recall curve (PR AUC).

The OpenBioLink benchmark dataset consists of 7 node and 30 edge types, covering a wide range of ontology terms, biomedical entities and their relationships (Fig. 1). Corresponding true negative edge types used in the dataset were either extracted directly from the data source or inferred from disjoint relation-type pairs (e.g. for gene–anatomy relationships, over-expression and under-expression data). Statistics about the dataset are available in the Supplementary Material. The benchmark dataset is available in four different quality filter settings (high, medium, low and all) which are based on confidence scores. These confidence scores are data source specific, corresponding thresholds for the different quality settings are taken from the documentation of the data sources. To be applicable to a wider variety of link prediction methods, the OpenBioLink benchmark graph is available in both a directed and an undirected version. In the undirected version each relationship is present only once in the dataset, whereas in the directed version additional explicit reverse edges for symmetric relations (e.g. 'interaction') are added. Licensing terms of integrated datasets are detailed in the documentation, and should be considered when redistributing the benchmark or any derivative work.

3 Discussion and future work

A preliminary baseline evaluation with the graph embedding methods TransE (Bordes *et al.*, 2013) and TransR (Lin *et al.*, 2015) was performed. Hyperparameter optimization was performed for each model, and the best model configuration was trained and tested against the OpenBioLink benchmark dataset. Details on hyperparameter estimation and per-relation results are available in Supplementary Tables S4 and S5. The best results with hits@10 of 7.5% over all relations were achieved by a TransE model with an embedding dimensionality of 100. This result reflects that established, simple graph embedding models can make some useful predictions on this benchmark, but there is still ample room for algorithmic improvement.

We will carry out more extensive evaluations, including other methods such as the metapath-based approach (Himmelstein *et al.*, 2017) and scalable rule learning (Meilicke *et al.*, 2019).

To further establish the OpenBioLink framework, we will host annual, public OpenBioLink benchmarking events so that a wide range of current and upcoming link prediction models can be evaluated, and the resources of the broader research community around link prediction can be better used for biomedical use cases. Future iterations of the benchmark dataset will be extended with additional knowledge from external resources such as Hetionet (Himmelstein *et al.*, 2017).

Eventually, predictions should be verified through experiments. Ultimately, they might help improve the generation of novel research hypotheses and become an important tool for driving the advancement of biomedical research.

Funding

This project received funding from the European Union's Horizon 2020 research and Innovation program under grant agreement no. 668353.

Conflict of Interest: none declared.

References

Ali,M. *et al.* (2019) BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. *Bioinformatics*, **35**, 3538–3540.

Alshahrani,M. *et al.* (2017) Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics*, **33**, 2723–2730.

Bordes,A. *et al.* (2013) Translating embeddings for modeling multi-relational data. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Curran Associates Inc., Red Hook, NY, USA, pp. 2787–2795.

Crichton,G. *et al.* (2017) Neural networks for link prediction in realistic biomedical graphs: a multidimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, **19**, 1–11.

Dettmers,T. *et al.* (2018) Convolutional 2D knowledge graph embeddings. In: McIlraith,S.A. *et al.* (eds) *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, CA, USA.

Dumontier,M. *et al.* (2014) Bio2rdf release 3: a larger connected network of linked data for the life sciences. In: Matthew,H. *et al.* (eds) *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, Vol. **1272**, CEUR-WS.org, Aachen, Germany, pp. 401–404.

Himmelstein,D.S. *et al.* (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**.

Lin,Y. *et al.* (2015) Learning entity and relation embeddings for knowledge graph completion. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Learning*, AAAI Press, Palo Alto, CA, US, pp. 2181–2187.

Meilicke,C. *et al.* (2019) Anytime bottom-up rule learning for knowledge graph completion. In: *International Joint Conferences on Artificial Intelligence Organization*, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* pp. 3137–3143.

Toutanova,K. and Chen,D. (2015) Observed versus latent features for knowledge base and text inference. In: Allauzen,A. *et al.* (eds) *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Association for Computational Linguistics, Beijing, China, pp. 57–66.

Yue,X. *et al.* (2020) Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, **36**, 1241–1251.