Knowledge Graph Embedding for Ecotoxicological Effect Prediction*

Erik B. Myklebust[1,2], Ernesto Jimenez-Ruiz[2,3], Jiaoyan Chen[4], Raoul Wolf[1], and Knut Erik Tollefsen[1]

[1] Norwegian Institute for Water Research (NIVA), Oslo, Norway
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Alan Turing Institute, London, United Kingdom
[4] Department of Computer Science, University of Oxford, United Kingdom

Abstract. Exploring the effects of a chemical compound on a species takes a considerable experimental effort. Appropriate methods for estimating and suggesting new effects can dramatically reduce the work needed to be done by a laboratory. Here, we explore the suitability of using a knowledge graph embedding approach for ecotoxicological effect prediction. A knowledge graph has been constructed from publicly available data sets, including a species taxonomy and chemical knowledge. These knowledge sources are integrated by ontology alignment techniques. Our experimental results show that the knowledge graph and its embeddings augment the baseline models.[1]

1 Introduction

It takes immense experimental efforts to determine ecotoxicological effects a chemical compound has on a species. These *effect* data is available for a narrow range of compound-species pairs and a limited number of experimental test.

Here, we present a preliminary study of the benefits of using Semantic Web tools to integrate different data sources and knowledge graph (KG) approaches to improve the ecotoxicological effect prediction over a baseline. Hence, our contribution is twofold:

(i) We have created a KG by gathering and integrating the relevant data from disparate sources. In order to discover equivalent entities we exploit internal resources, external resources (*e.g.*, Wikidata [16]) and ontology alignment (*e.g.*, LogMap [6, 5]).

(ii) We have evaluated three KG embedding approaches (TransE [2], DistMult [18] and HolE [12]) together with a baseline based on a one-hot encoding. Out evaluation shows improvement in the metrics using KG embedding for a majority of the selected classification models. Note that, recall is preferred over precision, *i.e.*, rather overestimate the effect of a chemical compound, than underestimate its hazardousness.

[1] This paper is a short version of our ISWC 2019 In-use paper [11].

2 Preliminaries

Knowledge graphs. We follow the RDF-based notion of KGs [1] which are composed by RDF triples $\langle s, p, o \rangle$, where s represents a subject (a class or an instance), p represents a predicate (a property) and o represents an object (a class, an instance or a data value *e.g.*, text, date and number).

Ontology alignment. Ontology alignment is the process of finding mappings or correspondences between a source and a target ontology or knowledge graph [3]. These mappings are typically represented as equivalences among the entities of the input resources (*e.g.*, `ncbi:DaphniaMagna` `owl:sameAs` `ecotox:daphniamagna`).

Embedding models. KG embedding [17] plays a key role in link prediction problems where the goal is to learn a scoring function $S : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$. $S(s, p, o)$ is proportional to the probability that a triple $\langle s, p, o \rangle$ is encoded as true. Several models has been proposed, *e.g.*, Translating embeddings model (TransE) [2]. These models are applied to KGs to resolve missing facts in largely connected KGs, such as DBpedia [9].

3 The TERA knowledge graph

We construct the *Toxicology and Risk Assessment* (TERA) KG from four sources: *(i)* The effect data is gathered from ECOTOX [15]. We focus our effort on acute effects, *e.g.*, LC50 (lethal concentration for 50% of test species) and NR-ZERO (no effect on all test species). This data is converted to a compound-species pair and a label (true or false). *(ii)* The chemical hierarchy is created by combining RDF data available from PubChem [8] and querying the ChEMBL [4] SPARQL endpoint. *(iii)* The species hierarchy is gathered from the tabular data available in the NCBI Taxonomy [14]. *(iv)* We gather species habitat and endemic data from the Encyclopedia of Life (EOL) [13]. We align the four data sources using LogMap and the Wikidata SPARQL endpoint. Details of the construction of the TERA knowledge graph is available in [10].

4 Effect prediction

We learn different types of classification models, including Gaussian naive-bayes (NB), quadratic discriminant analysis (QDA), radial basis function kernel support-vector machine (SVM), and multilayer perceptron (MLP), to solve the problem described in Figure 1. The input is a compound-species pair. It is encoded either as the the concatenation

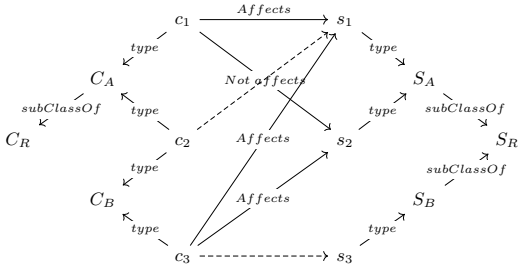
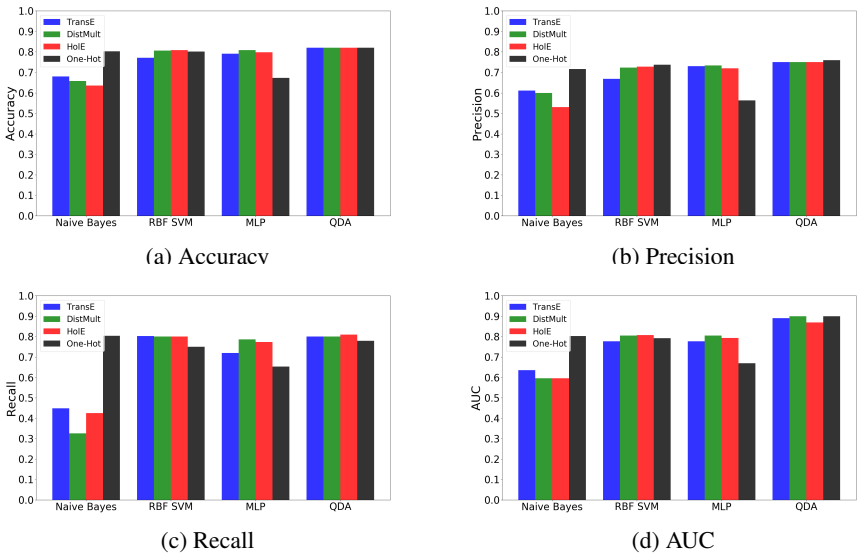


Fig. 1: The effect prediction problem. Lowercase s_j and c_i are instances of species and compounds, while uppercase denote classes in the hierarchy. Solid lines are observations and dashed lines are to be predicted. *i.e.*, does c_2 affect s_1?



(a) Accuracy (b) Precision

(c) Recall (d) AUC

Fig. 2: Prediction results for the different models.

of the one-hot vectors of the compound and the species (baseline), or the concatenation of the embedding vectors learned by the embedding model (TransE [2], DistMult [18] or HolE [12]). These models where considered since they are intuitive, have show state-of-the-art performance (*e.g.*, [7]), and encodes directional relationships, respectfully. The output is binary: *Affects* (1) and *Not affects* (0), representing the compound affects the species or not.

5 Results and Discussion

Results. Figure 2 shows the results of different models using different encoding methods of the input (compound-species pair). We find that two out of the four testing models, namely SVM and MLP achieve higher performance with KG embedding than with one-hot encoding. For the QDA model, KG embedding also has higher recall than one-hot encoding, although the overall metrics AUC and accuracy are similar. Note that recall is more important than precision in ecotoxicological effect prediction. The only exception is the NB model, where one-hot encoding has much higher performance than KB embedding. That is because NB holds the assumption that the input variables are conditional independent. Hence, it works better on the one-hot encoding which is quite sparse. However, it is worthwhile to note that the performance of NB with one-hot encoding does not outperform the MLP and QDA models with KB embedding.

Conclusion. We have created a KG called TERA that aims at covering the knowledge and data relevant to the ecotoxicological domain. We have also implemented a proof-of-concept prototype for ecotoxicological effect prediction based on knowledge graph embeddings and classification models. Some of the models used can take advantage of the learned embedded features. However, simple models like NB preferred the one-hot encoded vectors. The obtained results are encouraging, showing the positive impact

of using KG embedding models and the benefits of having an integrated view of the different knowledge and data sources.

Future work. The main goal in the long-term future is to make the TERA-KG accessible for domain researchers and improve the effect prediction by enriching the KG. In the near future, we intend to improve the current ecotoxicological effect prediction prototype and evaluate the suitability of more sophisticated models like Graph Convolutional Networks.

Resources. The datasets, evaluation results, documentation and source codes are available from the following GitHub repository: https://github.com/Erik-BM/NIVAUC

Acknowledgements. This work is supported by the grant 272414 from the Research Council of Norway (RCN), the MixRisk project (RCN 268294), the AIDA project (The Turing Institute) and the SIRIUS Centre for Scalable Data Access (RCN 237889).

References

1. Arnaout, H., Elbassuoni, S.: Effective Searching of RDF Knowledge Graphs. Web Semantics: Science, Services and Agents on the World Wide Web **48**(0) (2018)
2. Bordes, A., et al.: Translating Embeddings for Modeling Multi-relational Data. In: Advances in Neural Information Processing Systems 26, pp. 2787–2795 (2013)
3. Euzenat, J., Shvaiko, P.: Ontology Matching, Second Edition. Springer (2013)
4. Hastings, J., et al.: ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research **44**(D1), D12149 (January 2016)
5. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: Logic-Based and Scalable Ontology Matching. In: 10th International Semantic Web Conference. pp. 273–288 (2011)
6. Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: ECAI Conference. pp. 444–449 (2012)
7. Kadlec, R., Bajgar, O., Kleindienst, J.: Knowledge base completion: Baselines strike back. CoRR **abs/1705.10744** (2017), http://arxiv.org/abs/1705.10744
8. Kim, S., et al.: PubChem 2019 update: improved access to chemical data. Nucleic Acids Research **47**(D1), D1102–D1109 (10 2018)
9. Lehmann, J., et al.: DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2), 167–195 (2015)
10. Myklebust, E.B., Jiménez-Ruiz, E., Chen, J., Wolf, R., Tollefsen, K.E.: Enabling Semantic Data Access for Toxicological Risk Assessment. CoRR **abs/1908.10128** (2019)
11. Myklebust, E.B., Jiménez-Ruiz, E., Chen, J., Wolf, R., Tollefsen, K.E.: Knowledge graph embedding for ecotoxicological effect prediction. In: Int'l Sem. Web Conf. (ISWC) (2019)
12. Nickel, M., Rosasco, L., Poggio, T.A.: Holographic embeddings of knowledge graphs. CoRR **abs/1510.04935** (2015), http://arxiv.org/abs/1510.04935
13. Parr, C.S., et al.: The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. (2014)
14. Sayers, E.W., et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research **37**(suppl_1), D5–D15 (10 2008)
15. U.S. EPA: Ecotoxicology knowledgebase (ecotox) (2019), https://cfpub.epa.gov/ecotox/
16. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
17. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Trans. Knowl. Data Eng. **29**(12), 2724–2743 (2017)
18. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. CoRR **abs/1412.6575** (2015)