Semantically Enriched Datasets for Link Prediction: DB100k+, NELL-995+ and YAGO3-10+

Nicolas Robert, Pierre Monnin and Catherine Faron

Université Côte d'Azur, Inria, CNRS, I3S, Sophia-Antipolis, France
{nicolas.robert, pierre.monnin, catherine.faron}@inria.fr

Abstract. Knowledge graphs constitute a native neuro-symbolic experimental setting due to their logic foundations, which motivates the development of neuro-symbolic approaches for Link Prediction (LP). Since current LP reference datasets seldom involves ontological knowledge, benchmarking such approaches is difficult. That is why, starting from the widely accepted datasets DB100k, NELL-995 and YAGO3-10, we semantically enriched them with ontological knowledge, namely class hierarchy and relation signatures (domains and ranges), and inferred new entity type assertions to create DB100k+, NELL-995+ and YAGO3-10+. We also present a generic masking script to generate sub-graphs with variable proportions of triples with signed/partially signed (no domain or no range)/unsigned (no domain and no range) relations, to evaluate the impact of semantic information on learning performance.

1 Introduction

A *Knowledge Graph (KG)* is a set of triples $T \subseteq E \times R \times E$ where E is a set of entities and R a set of relations or predicates. Each triple (s, p, o) instantiates a relation between two entities (its *subject* and *object*), and represents a fact about a modeled world, *e.g.*, $(\text{Paris}, \text{capitalOf}, \text{France})$. Besides triples, KGs can also be associated with ontologies, i.e., a formal representations of the vocabulary used to model a domain. In such ontologies, axioms include, among others, entity typing with classes, the class hierarchy, and the signature (domain and/or range) of predicates. These specific ontology axioms will be referred in the remainder of this document as **semantic information**. We leave the consideration of other axioms for future work.

KGs are inherently incomplete because of their semi-automatic construction process, and this flaw is to be considered all along their life cycle [5]. Link Prediction (LP) aims at tackling this incompleteness by predicting either an object given $(s, p, ?)$ or a subject given $(?, p, o)$. KG Embedding Models (KGEMs) have been extensively used in LP [1]: the embedding vectors of the subject, predicate and object of a triple are used in a scoring function that outputs a plausibility score for the triple. To evaluate LP approaches, there are well-known and widely used KG-based datasets, among which *DB100k* [4], *NELL995* [10] and *YAGO3-10* [8]. They are subsets of existing KGs and only contain data assertions (i.e., factual triples). In particular, they do not integrate the ontology axioms that exist in such KGs, which prevents their usage in neuro-symbolic approaches. Yet, several recent works propose to exploit ontological knowledge in LP, either to provide new relevant evaluation metrics [6], to inject this knowledge into loss functions during training [7, 2], or to design KGEMs that inherently model this knowledge [9]. This growing collection of neuro-symbolic approaches thus emphasizes the need for benchmark datasets integrating ontological axioms to support their evaluation [3].

To fill this gap, we propose three new datasets, namely **DB100k+**, **NELL-955+** and **YAGO3-10+**, which extend the original well-known LP datasets with their related semantic information. We also present a generic masking script to generate sub-graphs with variable proportions of triples with fully signed (domain and range)/partially signed (no domain or no range)/unsigned (no domain and no range) relations, in order to evaluate the impact of semantic information availability on learning performance. The datasets, the source code to create them and the masking script are available online under the LGPL2.1 License on GitHub[1] and Zenodo[2].

The rest of this paper is organized as follows: Sections 2-4 respectively present NELL-995+, YAGO3-10+ and DB100k+; Section 5 presents the masking algorithm; Section 6 concludes.

2 NELL-995+

Never Ending Language Learning (NELL)[3] is a system running continuously since 2010 to extract facts from text of millions of web pages and to improve its own reading competence over time. The *NELL-995* [10] dataset is a subset of the 995th iteration of the NELL system[4]. It comprises 75,492 entities and 200 predicates occurring within the 149,678 triples in the train set, 543 triples in the validation set and 3,992 triples in the test set.

NELL-995+ is the semantically enriched version of *NELL-995* that we constructed using the original 995th iteration of NELL and the related 995th iteration of NELL ontology[5]. All of the 75,492 entities in the original dataset *NELL-995* have declared types. We removed type assertions using the general class `concept:everypromotedthing` to avoid too general and thus potentially useless statements. We added 350,723 new entity type assertions inferred based on the predicate domains and ranges and the class hierarchy provided in the NELL ontology.

[1] https://github.com/Wimmics/semantically-enriched-link-prediction-datasets
[2] https://doi.org/10.5281/zenodo.15834518
[3] https://nell-ld.telecom-st-etienne.fr/
[4] http://rtw.ml.cmu.edu/resources/results/08m/NELL.08m.995.esv.csv.gz
[5] http://rtw.ml.cmu.edu/resources/results/08m/NELL.08m.995.ontology.csv.gz

Table 1. Statistics of DB100k+, NELL-955+ and YAGO3-10+, in terms of number of fully signed, domain-only, range-only or unsigned relations, and their associated triples.

Dataset		# Fully signed	# Domain-only	# Range-only	# Unsigned	Total
DB100k+	Relations	278	76	83	33	470
	Triples	229,831	48,220	346,945	72,576	652,572
NELL995+	Relations	190	0	0	10	200
	Triples	114,335	0	0	0	39,878
YAGO3-10+	Relations	28	9	0	0	37
	Triples	1,067,137	21,903	0	0	1,089,040

3 YAGO3-10+

The Yet Another Great Ontology (YAGO) Knowledge Base[6] is a KG automatically built from Wikipedia, WordNet and GeoNames. *YAGO3-10* [8] is a subset of the 3rd version of the YAGO knowledge base[7]. It comprises 123,182 entities and 37 predicates occurring within 1,079,040 triples in the train set, 5,000 triples in the validation set, and 5,000 triples in the test set.

YAGO3-10+ is the semantically enriched version of *YAGO3-10* that we constructed as follows. We considered the entity type assertions in the `yagoTypes.ttl` file from the YAGO3 archive[8]. Out of the 123,182 entities occurring in this file, 121,009 have declared types (and 2,173 are untyped). We inferred 1,806,190 new entity type assertions based on the class hierarchy in the `yagoTaxonomy.ttl` file and relation domains and ranges in the `yagoSchema.ttl` file. Out of the 37 relations occurring in the assertions, 28 have both a domain and a range declared in YAGO schema and 9 relations have only a domain and no range.

4 DB100k+

DBpedia[9] is a cross-domain crowdsourced KG built from the information contained in the *infobox* of each Wikipedia article. *DB100k* [4] is a subset of the October 2016 version *DBpedia*[10]. This dataset uses 99,604 entities and 470 predicates in 597,572 triples in the train set, and 50,000 in both the validation and the test sets.

DB100k+ is the semantically enriched version of *DB100k* that we constructed as follows. Entities in *DB100k* follow the Wikidata notation while entity types in the DBpedia dump follow the DBpedia notation[11]. Hence, we linked the *DB100k* entities to DBpedia entities using the archived interlanguage alignment[12]. This processing revealed that, out of the 99,604 entities in *DB100k*, 92,558 have a type and that 8 entities are obviously not Wikidata entities (e.g. `index.html` or `?autoplay=true`). For the sake of compatibility with the original *DB100k* dataset and to avoid the risk of introducing wrong or noisy information, these entities were left untyped. In order to reduce noise, only entity type assertions using classes within the DBpedia ontology namespace were kept (e.g. assertions using class `owl:Thing` were removed). For semantic enrichment, predicate domains and ranges were obtained from the DBpedia ontology version of Oct. 2016[13]). Out of the 470 relations used in *DB100k*, 278 have a declared domain and range, 33 have neither domain nor range, 76 have a declared domain and no range, and 83 have a declared range and no domain. We added 216,738 new entity type assertions to *DB100k* by inferring types based solely on the class hierarchy. We decided not to infer types based on predicate domains and ranges since they produce noisy data on *DB100k* (e.g. England of type Country, but also Horse Race and Music Genre).

5 Masking algorithm

We designed and developed a generic masking algorithm to generate variations of the semantically enriched datasets, considering input target proportions of triples with signed, partially signed or unsigned relations. It is a greedy algorithm that: (1) represents each solution as a dictionary where each key is a predicate, and each value is a dictionary with two 'domain' and 'range' boolean keys (true meaning retained, false meaning masked); (2) evaluates the fitness of a solution as the sum of the differences between the target proportions and the obtained proportions in the train, validation set, and test sets; (3) aims at improving the solution at each iteration by masking a domain or a range such that the fitness is increased for the best; (4) stops when a solution cannot be enhanced this way.

To illustrate, we created a variant of NELL-995+ targeting the following proportions: 10% of triples with fully signed relations (a domain and a range), 30% with relations having a domain and no range, 10% with relations having a range and no domain, and 50% with unsigned relations. Our algorithm removed 180 of the 380 domain or range declarations and achieved the following triple proportions: 9,97%, 29,99%, 10,00% and 50,04%.

6 Conclusion

We created three KG-based datasets DB100k+, NELL-995+ and YAGO3-10+, by semantically enriching link prediction reference datasets. Table 1 summarizes the statistics of the number of relations and triples in each dataset, with additional statistics per split available on GitHub and Zenodo. We also provide a masking script to create subgraphs with variable proportions of triples with signed, partially signed, or unsigned relations. By extending LP reference datasets with ontological knowledge, these datasets will support the development and evaluation of neuro-symbolic approaches that take into account such knowledge, and ease their comparison with other approaches. For instance, as future work, we will use these datasets to evaluate the impact of type/domain/range information on LP performance.

[6] https://yago-knowledge.org/
[7] https://yago-knowledge.org/downloads/yago-3
[8] https://yago-knowledge.org/data/yago3/yago-3.0.2-turtle-simple.7z
[9] https://www.dbpedia.org/
[10] https://downloads.dbpedia.org/2016-10/core/
[11] https://downloads.dbpedia.org/2016-10/core/instance_types_en.ttl.bz2
[12] https://downloads.dbpedia.org/2016-10/core/interlanguage_links_chapters_en.ttl.bz2
[13] https://downloads.dbpedia.org/2016-10/dbpedia_2016-10.owl

References

[1] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, M. Galkin, S. Sharifzadeh, A. Fischer, V. Tresp, and J. Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, 44(12):8825–8845, 2022. doi: 10.1109/TPAMI.2021.3124805.

[2] C. d'Amato, N. F. Quatraro, and N. Fanizzi. Injecting background knowledge into embedding models for predictive tasks on knowledge graphs. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings*, volume 12731 of *Lecture Notes in Computer Science*, pages 441–457. Springer, 2021. doi: 10.1007/978-3-030-77385-4_26.

[3] C. d'Amato, L. Mahon, P. Monnin, and G. Stamou. Machine learning and knowledge graphs: Existing gaps and future research challenges. *Transactions on Graph Data and Knowledge*, 1(1):8:1–8:35, 2023. doi: 10.4230/TGDK.1.1.8.

[4] B. Ding, Q. Wang, B. Wang, and L. Guo. Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 110–121. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1011.

[5] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers, 2021. ISBN 978-3-031-00790-3. doi: 10.2200/S01125ED1V01Y202109DSK022.

[6] N. Hubert, P. Monnin, A. Brun, and D. Monticolo. Sem@K: Is my knowledge graph embedding model semantic-aware? *Semantic Web*, 14(6):1273–1309, 2023. doi: 10.3233/SW-233508.

[7] N. Hubert, P. Monnin, A. Brun, and D. Monticolo. Treat different negatives differently: Enriching loss functions with domain and range constraints for link prediction. In *The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I*, volume 14664 of *Lecture Notes in Computer Science*, pages 22–40. Springer, 2024. doi: 10.1007/978-3-031-60626-7_2.

[8] F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *Seventh Biennial Conference on Innovative Data Systems Research, CIDR 2015, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org, 2015. URL http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf.

[9] B. Xiong, N. Potyka, T. Tran, M. Nayyeri, and S. Staab. Faithful embeddings for $\mathcal{EL}^{++}$ knowledge bases. In *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, volume 13489 of *Lecture Notes in Computer Science*, pages 22–38. Springer, 2022. doi: 10.1007/978-3-031-19433-7_2.

[10] W. Xiong, T. Hoang, and W. Y. Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 564–573. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1060.