



HAL
open science

Schema First! Learn Versatile Knowledge Graph Embeddings by Capturing Semantics with MASCHInE

Nicolas Hubert, Heiko Paulheim, Pierre Monnin, Armelle Brun, Davy Monticolo

► To cite this version:

Nicolas Hubert, Heiko Paulheim, Pierre Monnin, Armelle Brun, Davy Monticolo. Schema First! Learn Versatile Knowledge Graph Embeddings by Capturing Semantics with MASCHInE. K-CAP 2023 - Knowledge Capture Conference 2023, Dec 2023, Pensacola, United States. pp.188-196, 10.1145/3587259.3627550 . hal-04344873

HAL Id: hal-04344873

<https://inria.hal.science/hal-04344873v1>

Submitted on 14 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Schema First! Learn Versatile Knowledge Graph Embeddings by Capturing Semantics with MASCHInE

Nicolas Hubert
Université de Lorraine, ERPI,
Université de Lorraine, CNRS, LORIA
Nancy, France
nicolas.hubert@univ-lorraine.fr

Heiko Paulheim
University of Mannheim
Mannheim, Germany
heiko.paulheim@uni-mannheim.de

Pierre Monnin
Univ. Côte d’Azur, Inria, CNRS, I3S
Sophia Antipolis, France
pierre.monnin@inria.fr

Armelle Brun
Université de Lorraine, CNRS, LORIA
Nancy, France
armelle.brun@loria.fr

Davy Monticolo
Université de Lorraine, ERPI
Nancy, France
davy.monticolo@univ-lorraine.fr

ABSTRACT

Knowledge graph embedding models (KGEMs) have gained considerable traction in recent years. These models learn a vector representation of knowledge graph entities and relations, a.k.a. knowledge graph embeddings (KGEs). Learning versatile KGEs is desirable as it makes them useful for a broad range of tasks. However, KGEMs are usually trained for a specific task, which makes their embeddings task-dependent. In parallel, the widespread assumption that KGEMs actually create a semantic representation of the underlying entities and relations (e.g., project similar entities closer than dissimilar ones) has been challenged. In this work, we design heuristics for generating *protographs* – small, modified versions of a KG that leverage RDF/S information. The learnt protograph-based embeddings are meant to encapsulate the semantics of a KG, and can be leveraged in learning KGEs that, in turn, also better capture semantics. Extensive experiments on various evaluation benchmarks demonstrate the soundness of this approach, which we call **Modular and Agnostic SCHEMA-based Integration of protograph Embeddings (MASCHInE)**. In particular, MASCHInE helps produce more versatile KGEs that yield substantially better performance for entity clustering and node classification tasks. For link prediction, using MASCHInE substantially increases the number of semantically valid predictions with equivalent rank-based performance.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence; Knowledge representation and reasoning*; • **Information systems** → *World Wide Web*.

KEYWORDS

Knowledge Graph Embeddings, Schema-based Representation Learning, Link Prediction, Entity Clustering, Node Classification

1 INTRODUCTION

Knowledge Graphs (KGs) such as DBpedia and YAGO represent facts as triples (h, r, t) formed by a head h and a tail t linked by a semantic relation r which qualifies the nature of their relationship. Typical learning problems with KGs are Link Prediction (LP), Entity Clustering (EC), and Node Classification (NC).

In most cases, these problems are addressed using Knowledge graph embedding models (KGEMs), which generate vector representations for entities and relations of a KG, a.k.a. Knowledge Graph Embeddings (KGEs). While generating dense and numerical vectors for entities and relations, KGEMs are expected to produce KGEs that retain the underlying semantics of the KG [21]. This widespread assumption that KGEMs create a semantic representation of the underlying entities and relations (i.e., project similar entities closer than dissimilar ones) has been challenged recently [14]. Their semantic capabilities presumably inherited from the rich information contained in the KG is neither fully satisfying nor consistent for tasks such as concept clustering [14]. In this work, we claim that this is because schemas (e.g. RDF/S and OWL) are often overlooked as an interesting source of information for improving embeddings.

Injecting schema-based information into the training phase could help enhancing the semantic awareness of KGEMs [13]. This way, better performance can be reached using available ontological information, without introducing major overhead in models’ complexity. A few works incorporate such schemas when training KGEM, whether it is in the loss function [7, 13], in the negative sampling procedure [15, 16], or in the model representations [6, 23]. All of those methods are usually bound to just one KGEM and therefore not universally applicable.

Another line of research that is less explored focuses on preprocessing graphs prior to encoding their components [5, 18], rather than focusing on the representation capability of the KGEM itself. This idea of summarizing [4, 9] or coarsening [18] RDF¹ graphs proves useful in many circumstances. The intended goals and expected benefits are multiple: scaling up training [10], initializing embeddings with more robust vector representations learnt on the coarsened graph [18], and improving predictive performance in KG-based downstream tasks [22]. It is worth a reminder that several heuristics exist for coarsening or summarizing graphs [4]. All these approaches, however, use statistics over the graphs, instead of considering schema information. In contrast, our work focuses on using RDF/S information to generate an abstraction of the KG that encapsulates general knowledge about how entities and relations interact based on entity types, and relation domains and ranges. We call such an abstraction a *protograph* (see left part of Figure 2).

¹Resource Description Framework. See <https://www.w3.org/RDF>

If considering schema-based information is usually done with the goal of improving results w.r.t. a given task such as LP [7, 13], it would be worth investigating whether the learnt embeddings actually have stronger semantic capabilities [13, 14]. If so, they are expected to provide better performance with regards to other tasks such as entity classification and clustering. In this work, we propose an approach named MASCHInE that leverages RDF/S information in a model-agnostic way, making it applicable to arbitrary KGEMs. Specifically, we create and embed *protographs* from a KG schema. The protograph embeddings are then used to initialize the actual KGEs, and training starts on these pre-trained KGEs. Finally, we assess the versatility of these KGEs by comparing results with a vanilla, traditional learning approach.

The main contributions of our work are summarized as follows:

- We devise two heuristics for building protographs that aim at encapsulating concise and meaningful information derived from the schema that underpins a KG.
- To the best of our knowledge, we propose the first work that studies the potential of pre-training embeddings with a protograph-assisted approach based on a schema, and subsequently use these embeddings for multiple tasks, e.g. link prediction, entity clustering and node classification.
- Through extensive experiments on several benchmarks, we demonstrate the value of using schema-based protographs as a means to generate versatile, multi-purpose embeddings.

The remainder of the paper is structured as follows. Related work is presented in Section 2. In Section 3, we detail our approach for building protographs and how they fit into our approach MASCHInE. In Section 4, the potential usefulness of embeddings learnt with MASCHInE is discussed, and experiments w.r.t. several benchmarks are carried out. Lastly, Section 5 sums up the main findings and outlines promising future research directions.

2 RELATED WORK

Knowledge graph embeddings. KGEMs have gained significant attention in recent years due to their ability to represent structured knowledge in a continuous vector space. The seminal translational model TransE [3] represents entities and relations as low-dimensional vectors and defines the relationship between a head, a relation, and a tail using a translation operation in the embedding space. Most of the KGEMs subsequently proposed aim at addressing its limitations and improve the representation expressiveness of KGEs [21], such as DistMult [24], ComplEx [20], ConvE [8], and TuckER [1]. Embeddings learnt by these KGEMs demonstrated potential applicability in tasks such as LP [21].

Schema-based representation learning. Many knowledge graphs are backed by a more or less expressive schema defining which classes of entities exist and by which relations they are allowed to interact, among others. Recent approaches in KG-based learning leverage the schema as a supplementary source of information to learn higher-quality embeddings for the task at hand. A common approach is to alter the negative sampling procedure in order to produce more realistic negative triples. For instance, type-constrained negative sampling (TCNS) [16] replaces the head or the tail with a random entity belonging to the same type as the ground-truth

entity. The model itself can be adapted to take schema-based information into account. In TaRP [6], type and entity-level information are simultaneously considered and encoded as prior probabilities and likelihoods of relations, respectively. Altering the loss function with schema-based information is another line of research. In [7], background ontological information is injected in the pairwise hinge loss function as constraints. Hubert *et al.* [13] leverage domains and ranges of relations, and propose semantic-driven loss functions that extend the most frequently used loss functions in LP.

3 PROPOSED APPROACH

In this section, we detail the two heuristics we propose for building protographs using schema-based information. Next, we provide some perspective on the overall proposed approach MASCHInE.

3.1 Protographs

Most KGs are underpinned by a schema, e.g., DBpedia, Wikidata, and YAGO. Generating a protograph using schema-based information to improve KGEM performance with respect to a given task is an under-explored avenue.

The protographs we build leverage RDF/S information and contain an entity p_C for each class C , and the same set of relations as the KG. The first design principle consists in adding an edge for each relation with domain and range restrictions, i.e., given the two axioms $domain(r, C_i)$, $range(r, C_j)$, we add a triple (p_{C_i}, r, p_{C_j}) to the protograph². We refer to this strategy as **P1**.

The protograph P1 contains as many triples as there are relations with both a domain and range³ in the KG. This may lead to relatively small-sized protographs from which there is little data for the KGEM to learn from. In addition, relation domains and ranges can refer to generic classes. If entities are typed in a fine-grained way – i.e. with more specific classes – such classes will be absent from the protograph. As such, we propose a second design principle. Assuming we observe the following axioms $domain(r, C_i)$, $range(r, C_j)$, $subclassOf(C'_i, C_i)$, $subclassOf(C'_j, C_j)$, we add the following triples to the protograph: (p_{C_i}, r, p_{C_j}) , $(p_{C'_i}, r, p_{C_j})$, $(p_{C_i}, r, p_{C'_j})$. In other words: for each subclass of the class appearing in the domain or range of a relation, an additional triple is created.

We refer to this protograph as **P2**. Figure 1 illustrates this process for a triple extracted from DBpedia.

P2 usually contains more triples than P1, at the possible expense of containing dubious ones. To illustrate, the DBpedia ontology contains the triples: $domain(currentWorldChampion, Sport)$, $range(currentWorldChampion, Agent)$, and $subclassOf(Family, Agent)$. They lead to the triple $(P_{Sport}, currentWorldChampion, P_{Family})$ for which the relevancy and correctness are questionable. Conceptually, with P2, we accept some noise to be introduced in the protograph for the benefit of more triples being generated. However, the generation of dubious triples is limited by two factors: firstly, when adding triples that derive from an original triple (i.e. containing the domain and range of the relation), we only consider direct subclasses – in contrast to transitive subclasses. Secondly, we fix one side of the triple and generate as many new triples as there

²Note that C_i and C_j might be the same, so there can be cycles in the protograph.

³In the KGs used in this paper, all relations have an explicit domain and range.

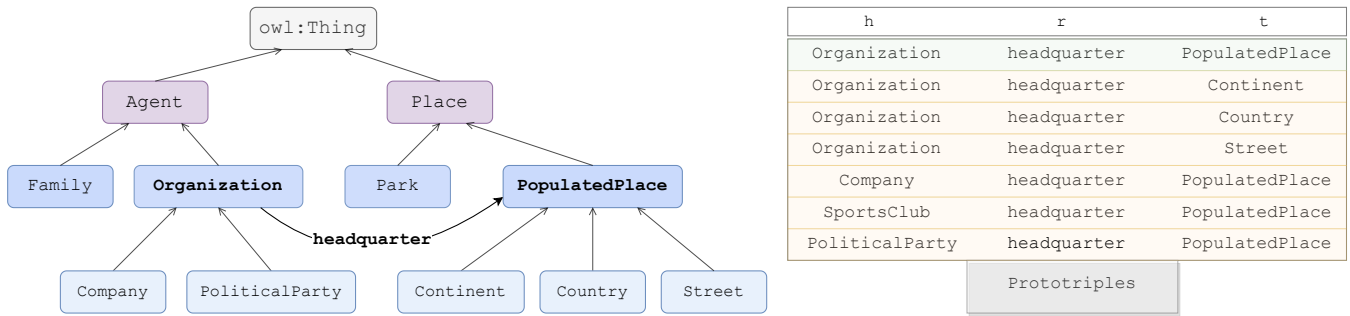


Figure 1: Excerpt from DBpedia class hierarchy (left) and the generated prototriples (right) according to P2. The full URI names have been shortened due to space limitations. The prototriple highlighted in green (right) comes from the schema, whereas the following triples are generated by fixing the domain (resp. range) of `dbo:headquarter` and replacing the class in the range (resp. domain) with each of its direct subclasses.

are direct subclasses of the class of the other side, i.e., we avoid generating triples (pC_i, r, pC'_i) .

3.2 MASCHInE

The proposed MASCHInE approach allows for generating pre-trained embeddings based on a protograph, to be further fine-tuned on the initial KG. MASCHInE is intended to produce more versatile KGEs that can be successfully used in various tasks. Conceptually, MASCHInE is made of four different modules that are depicted in Figure 2 and further detailed below.

(a) **Protograph building** is the preliminary step that generates a second source of data to learn from. In particular, schema-based information is leveraged to create a protograph that encapsulates more general knowledge about how entities and relations interact based on ontological constraints such as entity types, relation domains and ranges, etc. A mapping dictionary is created to pair entities between the KG and the protograph (the set of relations in the KG and the protograph are identical). In this work, each KG entity maps to its (potentially multiple) most specific classes – represented as entities in the protograph. In Section 3.1, we presented two heuristics for building protographs with such constraints. Other possibilities exist and we leave them for future work.

(b) **Protograph KGE learning** is the task of learning KGEs on the protograph itself. This can be achieved regardless of the KGEM at hand and is therefore model-agnostic⁴.

(c) **Instance KGE initialization** consists in transferring protograph embeddings to the corresponding entity and relation embeddings of the KG. To do so, we assign each entity in the KG the embedding vector of its corresponding protograph entity using the dictionary created in step (a) (i.e., its most specific class). When an entity maps to multiple protograph entities (i.e. when there exist multiple most specific classes for a given entity), the protograph embeddings of its most specific classes are averaged and the resulting embedding is transferred to the entity.

(d) **Fine-tuning** is performed on the entity KGEs. At this stage, embeddings learnt on the protograph could still intervene in the

fine-tuning procedure. In this work, however, we stick to plain entity KGE initialization, i.e. protograph embeddings are frozen after being transferred, and do not interact with entity KGEs anymore.

4 EXPERIMENTS

In our experiments, we apply the MASCHInE approach presented in Section 3. We experiment with five popular KGEMs. We first train on protographs, and then use the protograph embeddings to initialize entity and relation embeddings in the KG. Training starts on the KG and the fine-tuned KGEs are ultimately assessed in three tasks, namely LP, EC, and NC. For each task, we compare the vanilla setting (V), i.e., directly training the respective KGEM on the KG, with the MASCHInE approach based on P1 (resp. P2). Datasets, optimal hyperparameters, scripts, and pre-trained embeddings are publicly released⁵. In the following, protographs, KG entities and relations are encoded using five mainstream KGEMs: TransE [3], DistMult [24], ComplEx [20], ConvE [8], and Tucker [1].

4.1 Link Prediction

Table 1: KG and protograph characteristics. Column headers from left to right: number of entities, relations, and triples.

Dataset		$ \mathcal{E} $	$ \mathcal{R} $	$ \mathcal{T} $
YAGO14k	KG	14, 178	37	19, 183
	P1	22	37	37
	P2	590	37	4, 959
FB15k187	KG	14, 305	187	278, 436
	P1	138	187	187
	P2	138	187	187
DBpedia77k	KG	76, 651	150	190, 028
	P1	55	150	150
	P2	186	150	3, 210

Datasets. Due to the need for schema-defined KGs, the following three benchmark datasets are considered: YAGO14k, FB15k187, and DBpedia77k [13]. In particular, all entities are typed and relations have both a defined domain and range. Table 1 provides statistics for

⁴In our experiments, the same KGEM is used for both pre-training over the protograph and fine-tuning embeddings on the KG.

⁵<https://github.com/nicolas-hbt/versatile-embeddings>

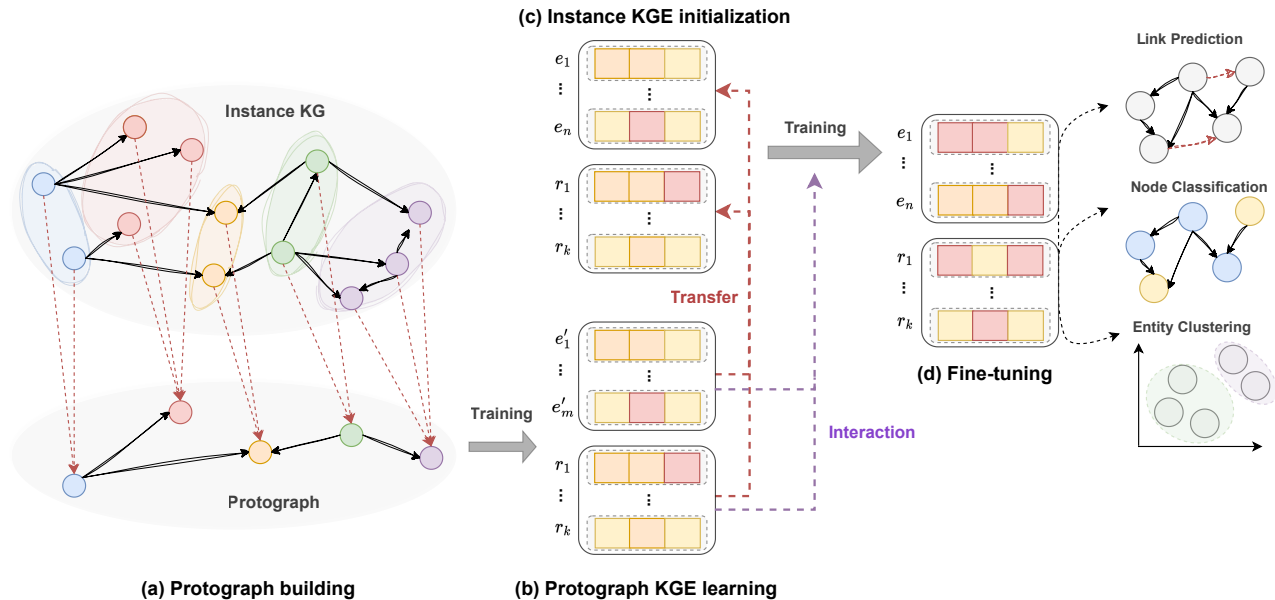


Figure 2: Overview of MASCHInE. The overall training approach is based on several modules that each features different possibilities. First, the protograph is built following predefined design principles (a). Protograph embeddings are learnt (b) before being transferred to the corresponding entities and relations of the KG (c). Next, training starts on the KG (d). During this step, embeddings learnt on the protograph can still interact in the fine-tuning procedure (purple dotted arrow). Finally, the fine-tuned KGEs can be used in several tasks.

Table 2: Link prediction results on YAGO14k, FB15k187, and DBpedia77k using KGEMs trained without protograph (V), with P1, and P2. For a given dataset and KGEM, comparisons are made between the three settings, and best results are in bold fonts. Hits@K and Sem@K are abbreviated as H@K and S@K, respectively.

		YAGO14k					FB15k187					DBpedia77k				
		MRR	H@3	H@10	S@3	S@10	MRR	H@3	H@10	S@3	S@10	MRR	H@3	H@10	S@3	S@10
TransE	V	0.766	0.806	0.885	0.987	0.971	0.232	0.259	0.430	0.970	0.958	0.236	0.285	0.405	0.983	0.976
	P1	0.746	0.791	0.871	0.994	0.993	0.227	0.254	0.425	0.970	0.959	0.217	0.260	0.382	0.985	0.980
	P2	0.700	0.762	0.854	1.000	0.999	0.227	0.254	0.425	0.970	0.959	0.216	0.260	0.380	0.987	0.982
DistMult	V	0.899	0.916	0.926	0.658	0.490	0.228	0.256	0.412	0.985	0.969	0.282	0.308	0.391	0.845	0.804
	P1	0.164	0.169	0.314	1.000	1.000	0.206	0.228	0.394	0.988	0.980	0.194	0.219	0.303	0.928	0.939
	P2	0.289	0.325	0.511	0.999	0.999	0.206	0.228	0.394	0.988	0.980	0.190	0.207	0.287	0.937	0.947
ComplEx	V	0.923	0.930	0.933	0.723	0.587	0.203	0.223	0.399	0.942	0.923	0.286	0.317	0.394	0.823	0.754
	P1	0.897	0.916	0.932	0.907	0.848	0.185	0.201	0.371	0.922	0.900	0.209	0.226	0.302	0.856	0.839
	P2	0.914	0.930	0.934	0.914	0.854	0.185	0.201	0.371	0.922	0.900	0.209	0.225	0.306	0.913	0.890
ConvE	V	0.934	0.936	0.940	0.858	0.824	0.265	0.297	0.463	0.979	0.971	0.227	0.259	0.352	0.897	0.889
	P1	0.927	0.930	0.935	0.882	0.845	0.272	0.301	0.470	0.977	0.970	0.227	0.256	0.352	0.932	0.932
	P2	0.931	0.931	0.938	0.935	0.922	0.272	0.301	0.470	0.977	0.971	0.229	0.259	0.356	0.943	0.942
TuckER	V	0.919	0.931	0.942	0.899	0.821	0.282	0.311	0.467	0.996	0.991	0.274	0.305	0.401	0.988	0.985
	P1	0.922	0.933	0.941	0.951	0.925	0.286	0.317	0.472	0.995	0.985	0.277	0.308	0.402	0.987	0.984
	P2	0.918	0.930	0.941	0.971	0.939	0.286	0.317	0.472	0.995	0.985	0.275	0.304	0.403	0.989	0.986

these datasets and their respective protographs P1 and P2. Note that protograph entities are classes from the original KG. Full dataset description is provided at: <https://github.com/nicolas-hbt/versatile-embeddings#datasets>.

Evaluation metrics. This section is motivated by evaluating the fine-tuned KGEs for LP. Performance is assessed w.r.t. Mean Reciprocal Rank (MRR), Hits@K, and Sem@K [11, 12]. The latter

accounts for the proportion of triples whose predicted head (resp. tail) belongs to the domain (resp. range) of the relation.

Implementation details. KGEMs were implemented in PyTorch. In accordance with [13], KGEMs are trained during 400 epochs. For TransE, DistMult, and ComplEx, uniform random negative sampling [3] was used to pair each train triple with one negative counterpart. ConvE and TuckER are trained using 1-N scoring [8]. In this work, we stick with the loss function originally used for

each KGEM. Evaluation is performed every 10 epochs and the best model w.r.t. MRR on the validation set is saved for performance assessment on the test set. The aforementioned procedure prevails for the vanilla training mode. For P1 and P2, we use the same procedure as for training on the KG, except that 200 epochs of training are previously performed on the protograph, for a total of 600 training epochs.

Experimental results. LP results are reported in Table 2. On YAGO14k and DBpedia77k, learning embeddings with P1 or P2 provides better Sem@K values compared to V. No significant change is noticed on FB15k187. This is likely to be an artefact of the Freebase class hierarchy and how entities are typed in FB15k187: the class hierarchy only has 2 levels, all relation domains and ranges are level-2 classes (therefore P1 and P2 are the same, see Table 1) and entities share many classes together. In this situation, the protograph may not contain discriminative enough information for pre-training embeddings, which is reflected in the similar results achieved regardless of the training setting. In addition, we observe that some KGEMs seem to benefit more from the MASCHInE approach. In particular, ConvE and TuckER tend to perform better w.r.t. both rank-based and semantic-oriented metrics. Overall, we notice that including protographs in the training phase helps making semantically valid predictions, as evidenced by the better Sem@K under P1 and P2. Protographs may help organize the embedding space in such a way that geometric distances between embeddings better reflect their semantic similarities. Regarding rank-based metrics, we notice equivalent performance, which could be explained since LP relies less on semantics and more on relational patterns [2] compared to other tasks such as entity clustering (see below).

4.2 Entity Clustering

Jain *et al.* [14] demonstrate that embeddings do not consistently capture the semantics of the KG (in particular, they only considered schema-agnostic approaches). In this section, we investigate whether embeddings learnt for the LP task under P1 and P2 can actually prove useful in EC.

4.2.1 Experiments on the original datasets. We investigate whether the embeddings learnt for the LP task (see Section 4.1) under P1 and P2 allow a better class separability for the task of EC.

Datasets. We reuse YAGO14k, FB15k187, and DBpedia77k as presented in Section 4.1. The ground-truth labels are the most-generic classes an entity belongs to. It should be noted that entities belonging to several most-generic classes are filtered out as they would be assigned multiple ground-truth labels.

Implementation details. Entity embeddings are retrieved at the best epoch on the validation sets of YAGO14k, FB15k187, and DBpedia77k, for all the KGEMs and under the three settings V, P1, and P2. Then, following the evaluation protocol in Jain *et al.* [14], the k-means clustering algorithm is run using default parameters of scikit-learn⁶.

Evaluation metrics. Clustering results are evaluated using the Adjusted Rand Index (ARI) – which measures the similarity between the cluster assignments by making pairwise comparisons –

and Normalized Mutual Information (NMI) – which measures the agreement between the cluster assignments. It should be noted that these metrics are applicable as we have the ground-truth labels for each entity. In particular, we compare clusters output by k-means with ground-truth labels that are the classes entities belong to. ARI and NMI results are reported in Figure 3. Lowest and highest scores are 0 (−1 for ARI) and 1, respectively.

Experimental results. Except for TuckER on FB15k187, both settings P1 and P2 provide substantial improvements. It seems that the entity embeddings learnt by TuckER are not able to cluster entities based on their classes, which diminishes the relevancy of using this model for drawing comparisons. That being said, the benefit of P1 and P2 is consistent across datasets. This is even more striking for YAGO14k, as evidenced in Figure 4: entities sharing the same ground-truth class tend to cluster more under P2. Jain *et al.* [14] pointed out the limitations of using KGEs for semantic representation of the underlying entities and relations. The relatively low performance of embeddings for EC under V (Figure 3 and Figure 4) indeed suggests that the semantic similarity between entities is not reflected through geometric similarities in their embeddings. However, P1 and P2 generate entity latent representations whose geometric similarities are more representative of their classes.

4.2.2 Experiments on GEval clustering problems. In this follow-up entity clustering experiment, the GEval evaluation framework [17] is used. GEval provides gold standard datasets based on DBpedia and suggest clustering models, configurations and evaluation metrics⁷.

Datasets. GEval proposes 4 datasets based on DBpedia. The first dataset contain entities of classes `dbo:AmericanFootballTeam` and `dbo:BasketballTeam`. We do not use this dataset as DBpedia77k does not contain any of these entities. Instead, we use the other three datasets for which DBpedia77k has a biggest coverage. For clarity, we name these datasets as CC, CCB, and CAMAP. The first two ones contain different numbers of DBpedia entities of classes `dbo:Cities` and `dbo:Countries`, while the third dataset contains five types of entities. Therefore, there are two ground-truth clusters in the first two datasets and five clusters in the third one.

Implementation details. We rely on GEval guidelines and perform the preliminary filtering step described as follows. First, we select entity embeddings learnt on DBpedia77k for the LP task (see Section 4.1). Embeddings of entities that do not appear in the GEval benchmark datasets are filtered out. Clustering with all the algorithms suggested in the GEval framework is performed on the remaining entities and evaluated w.r.t. several metrics.

Evaluation metrics. As in the previous experiment, we use ARI. Following the GEval guidelines, we additionally use Adjusted Mutual Information Score (AMI), V-Measure (VM), Fowlkes-Mallow Score (FM), Homogeneity (H), and Completeness (C). V-Measure is the harmonic mean of homogeneity and completeness measure, while the Fowlkes-Mallow Score is the geometric mean of pairwise precision and recall. All these metrics measure the correctness of the cluster assignments based on ground-truth labels. For each of these datasets, and under each setting, results achieved with the

⁶<https://scikit-learn.org/stable/index.html>

⁷<https://github.com/mariaangelapellegrino/Evaluation-Framework>

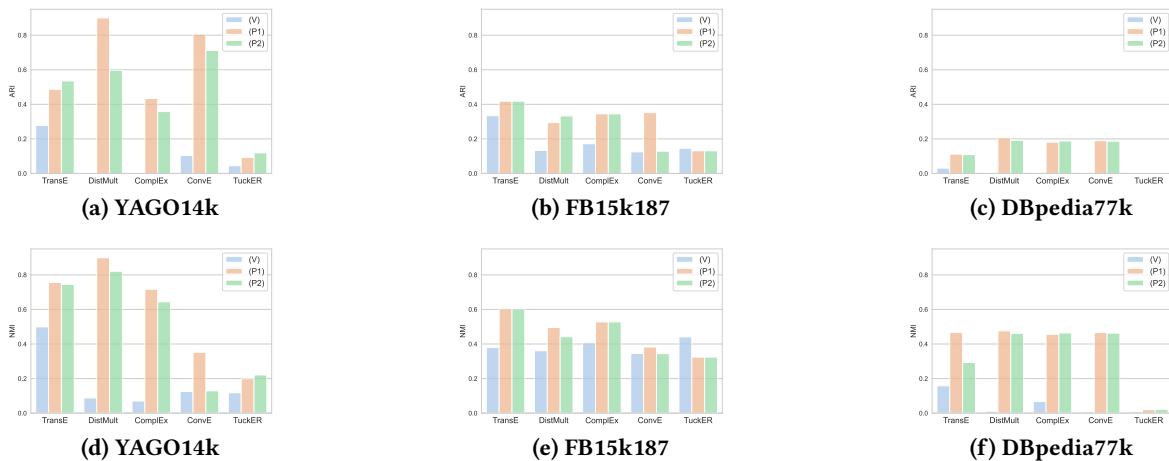


Figure 3: Entity clustering results on YAGO14k, FB15k187, and DBpedia77k.

Table 3: Entity clustering results on the GEval datasets. Bold font indicates which setting performs the best for a given model and dataset, while underlined results correspond to the second best setting.

		CC						CCB						CAMAP					
		ARI	AMI	VM	FM	H	C	ARI	AMI	VM	FM	H	C	ARI	AMI	VM	FM	H	C
TransE	V	0.179	0.140	0.141	0.596	0.143	0.139	0.183	0.147	0.147	0.649	0.164	0.133	0.564	0.783	0.785	0.736	0.951	0.668
	P1	<u>0.621</u>	<u>0.514</u>	<u>0.515</u>	<u>0.823</u>	<u>0.505</u>	<u>0.526</u>	<u>0.288</u>	<u>0.311</u>	<u>0.311</u>	<u>0.692</u>	<u>0.349</u>	<u>0.280</u>	<u>0.593</u>	<u>0.795</u>	<u>0.796</u>	<u>0.756</u>	<u>0.956</u>	<u>0.682</u>
	P2	0.680	0.570	0.571	0.849	0.563	0.579	0.299	0.314	0.314	0.698	0.353	0.284	0.646	0.815	0.816	0.791	0.964	0.708
DistMult	V	0.018	0.024	0.027	0.709	0.015	0.095	-0.024	0.009	0.010	0.770	0.006	0.028	-0.054	0.099	0.106	0.478	0.087	0.136
	P1	0.891	<u>0.817</u>	<u>0.817</u>	0.948	<u>0.813</u>	<u>0.821</u>	0.762	0.673	0.673	0.907	0.708	0.642	<u>0.655</u>	<u>0.705</u>	<u>0.707</u>	<u>0.798</u>	<u>0.782</u>	<u>0.646</u>
	P2	0.891	<u>0.824</u>	<u>0.824</u>	0.948	<u>0.818</u>	<u>0.830</u>	<u>0.756</u>	<u>0.626</u>	<u>0.627</u>	0.907	<u>0.646</u>	<u>0.608</u>	<u>0.933</u>	<u>0.872</u>	<u>0.872</u>	<u>0.963</u>	<u>0.886</u>	<u>0.859</u>
ComplEx	V	-0.011	0.009	0.011	0.694	0.007	0.033	-0.024	0.009	0.010	0.771	0.006	0.028	0.139	0.345	0.350	0.520	0.341	0.360
	P1	0.897	0.827	0.828	0.951	0.823	0.832	0.619	0.560	0.560	0.845	0.606	0.521	0.969	0.916	0.917	0.983	0.939	0.896
	P2	<u>0.319</u>	<u>0.264</u>	<u>0.265</u>	<u>0.665</u>	<u>0.269</u>	<u>0.262</u>	<u>0.493</u>	<u>0.460</u>	<u>0.460</u>	<u>0.788</u>	<u>0.507</u>	<u>0.422</u>	<u>0.961</u>	<u>0.914</u>	<u>0.914</u>	<u>0.978</u>	<u>0.938</u>	<u>0.892</u>
ConvE	V	0.311	0.234	0.235	0.691	0.223	0.249	0.271	0.242	0.243	0.687	0.271	0.220	0.591	0.608	0.611	0.76	0.672	0.559
	P1	<u>0.878</u>	<u>0.805</u>	<u>0.805</u>	<u>0.942</u>	<u>0.799</u>	<u>0.811</u>	<u>0.719</u>	<u>0.628</u>	<u>0.628</u>	<u>0.889</u>	<u>0.665</u>	<u>0.596</u>	<u>0.934</u>	<u>0.880</u>	<u>0.881</u>	<u>0.964</u>	<u>0.928</u>	<u>0.839</u>
	P2	0.916	0.848	0.848	0.960	0.847	0.849	0.790	0.697	0.697	0.919	0.727	0.669	0.954	0.915	0.915	0.975	0.968	0.868
TuckER	V	0.179	0.138	0.139	0.596	0.141	0.137	0.158	0.120	0.121	0.639	0.135	0.109	0.396	0.526	0.529	0.618	0.636	0.453
	P1	<u>0.069</u>	<u>0.051</u>	<u>0.052</u>	<u>0.543</u>	<u>0.053</u>	<u>0.051</u>	<u>0.137</u>	<u>0.094</u>	<u>0.094</u>	<u>0.631</u>	<u>0.105</u>	<u>0.086</u>	<u>0.376</u>	<u>0.635</u>	<u>0.637</u>	<u>0.599</u>	<u>0.795</u>	<u>0.531</u>
	P2	0.047	0.042	0.043	0.533	0.043	0.042	0.033	0.051	0.052	0.578	0.058	0.046	0.276	0.487	0.490	0.520	0.621	0.405

best performing clustering method are shown in Table 3. Lowest and highest scores are 0 (−1 for ARI) and 1, respectively.

Experimental results. First, it is worth noting that results on CAMAP are significantly higher than on the other two datasets. Recall that CAMAP includes entities belonging to five different classes, whereas CC and CCB only contain entities from `dbo:City`s and `dbo:Country`s. In this experiment, the embeddings learnt by the KGEMs are most useful for performing EC with five ground-truth clusters than with two, which might seem counterintuitive. However, in [14], it is demonstrated that in a KG, only a small subset of its entities actually have an embedding with a meaningful semantic representation. This is the reason why for the clustering task, results are not consistent across subsets of entities [14]. This is in line with our experimental findings that lead to the same conclusion.

[25] is also concerned with the ability of KGEMs to embed entities from the same class to the same region of the embedding space. Their results suggest that entities belonging to rare classes

have embeddings that are less separable w.r.t. to entity embeddings of other classes compared with entities belonging to frequent classes. This assumption needs to be qualified in light of our experimental findings: lots of entities in DBpedia77k are `dbo:City` and `dbo:Country`. Following [25], we should expect better results for CC and CCB in Table 3. However, it should be noted that both `dbo:City` and `dbo:Country` are subclasses of `dbo:Place` (equivalent to `dbo:Location`), which is over-represented in DBpedia77k⁸. It is arguably harder to separate entity embeddings for CC and CCB, as all of them actually belong to the same parent class. In contrast, CAMAP contain albums, cities, companies, movies, and universities. In this case, `dbo:City` is a child class of `dbo:Place`, `dbo:Company` and `dbo:University` are child classes of `dbo:Agent` while `dbo:Album` and `dbo:Movie` are child classes of `dbo:Work`.

⁸In DBpedia77k, 11, 586 entities belong to the `dbo:Place` class.

Table 4: Node classification results on the 12 DLCC test cases. Listed are the results of the best classifier for each combination of model and test case. Bold fonts indicate which setting performs the best for a given model and test case. Underlined results show which combination of model and setting performs the best for each test case.

Test Case	Setting	TransE	DistMult	Complex	ConvE	TuckER	Test Case	Setting	TransE	DistMult	Complex	ConvE	TuckER
tc01	V	0.702	0.810	0.820	0.885	0.832	tc02	V	0.654	0.564	0.566	0.732	0.669
	P1	0.647	0.754	0.784	0.875	0.857		P1	0.566	0.554	0.554	0.764	0.637
	P2	0.637	0.794	0.769	0.852	0.792		P2	0.571	0.627	0.627	0.687	0.679
tc03	V	0.584	0.571	0.554	0.629	0.669	tc04	V	0.638	0.540	0.512	0.818	0.570
	P1	0.589	0.541	0.519	0.699	0.664		P1	0.760	0.812	0.832	0.870	0.552
	P2	0.521	0.544	0.526	<u>0.622</u>	0.556		P2	0.805	0.802	0.818	0.818	0.680
tc05	V	0.704	0.683	0.678	0.822	0.814	tc06	V	0.645	0.940	0.938	0.940	0.950
	P1	0.887	0.937	0.940	0.940	0.814		P1	0.718	0.875	0.898	0.935	0.950
	P2	0.937	0.940	0.940	0.940	0.902		P2	0.740	0.828	0.788	0.895	0.942
tc07	V	0.578	0.515	0.552	0.560	0.522	tc08	V	0.552	0.552	0.520	0.598	0.575
	P1	0.505	0.540	0.560	0.560	0.552		P1	0.585	0.560	0.558	0.575	0.580
	P2	0.582	0.605	0.655	0.675	0.648		P2	0.595	0.588	0.588	0.578	0.672
tc09	V	0.552	0.522	0.525	0.730	0.745	tc10	V	0.578	0.572	0.562	0.640	0.690
	P1	0.560	0.538	0.552	0.745	0.780		P1	0.598	0.550	0.538	0.625	0.692
	P2	0.550	0.615	0.598	0.672	<u>0.742</u>		P2	0.538	0.510	0.535	0.590	0.588
tc11	V	0.632	0.518	0.535	0.688	0.700	tc12	V	0.650	0.538	0.552	0.702	0.708
	P1	0.628	0.542	0.522	0.618	0.685		P1	0.572	0.530	0.565	0.708	0.700
	P2	0.630	0.552	0.558	0.638	0.678		P2	0.625	0.802	0.738	0.802	0.752

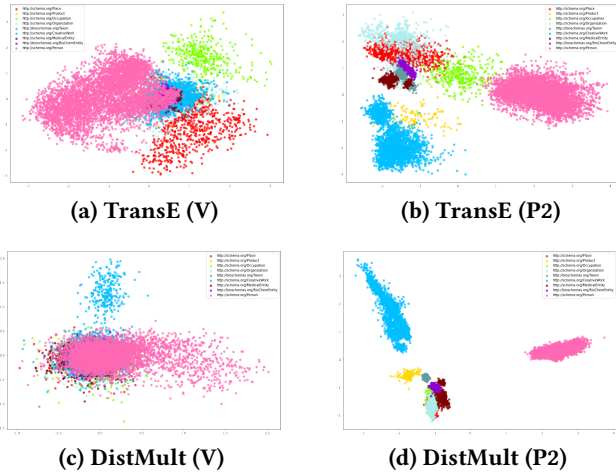


Figure 4: PCA visualizations of YAGO14k entity embeddings using the first two principal components. Each point represents an entity, whose color denotes its ground-truth class.

With three different generic classes, the semantic similarities between entities in CAMAP is more easily reflected into the geometric similarities of their embeddings.

Interestingly, the discrepancies in results, *i.e.*, better results on CAMAP compared with CC and CCB, are setting-dependent (V, P1, P2). Indeed, they are alleviated when using either P1 or P2. Under V, TransE, ComplEx, and ConvE provide substantially better results on CAMAP compared to CC. Once using P2, we cannot see any significant difference in results on CAMAP compared to CC for these models. This means that training with a protograph-assisted approach led the KGEMs to generate higher-quality embeddings in terms of class separability, especially for entities in CC and CCB that are harder to distinguish. With the sole exception of TuckER

whose performance may be due to non-optimal hyperparameter tuning, strong improvements are achieved w.r.t. the EC task on all datasets. Interestingly, not only V is outperformed by either P1 or P2 in each scenario, but actually by both of them at the same time. As a result, our schema-based, protograph-assisted learning approach MASCHInE proves effective for EC, regardless of the protograph design principle.

4.3 Node Classification

Following on from the previous clustering experiment that clearly showed the benefit of incorporating pre-trained protograph embeddings, in this section we are interested in performing node classification with such embeddings. In this experiment, entities are labelled on the basis of whether they comply with some description logic (DL) constructors.

Datasets. In this experiment, we use the synthetic DLCC node classification benchmark [19]. The DLCC benchmark helps analyze the quality of KGEs by evaluating them w.r.t. which kinds of DL constructors they can represent. We use the synthetic gold standard datasets⁹. These are 12 synthetic datasets featuring labelled entities to classify and underpinned by a schema defining entity types, relation domains and ranges. The synthetic part of DLCC has been chosen because the authors argue that those pose classification problems which are not influenced by any side effects and correlations in the knowledge graph.

Implementation details. For these 12 test cases, their respective protographs are built. Then, the KGEMs are trained w.r.t. the LP task under the three different settings V, P1, and P2. In line with [19], multiple classifiers were trained for each test case: decision trees (DT), naive bayes (NB), k-nearest neighbors (k-NN), support vector machine (SVM), and multilayer perceptron (MLP). Based on the

⁹<https://github.com/janathan/DL-TC-Generator>

best entity embeddings learnt during LP, these models are used to classify entities using the default scikit-learn parameters.

Evaluation metrics. NC performance is measured using F-score, which is the harmonic mean of precision and recall. Lowest and highest scores are 0 and 1, respectively. For each model and test case, results achieved with the best classifier are reported in Table 4.

Experimental results. First, not all of the gold standard test cases are equally hard for the task of NC. For instance, entities in tc01, tc05, and tc06 are obviously easier to label than in the other datasets. Particularly, the DLCC synthetic gold standard datasets are built on the basis of different DL constructors in order to analyze which KG constructs are more easily recognized by KGEMs.

For the less trivial cases, the gains that can be achieved by using protographs are more significant. Especially in cases which are fairly badly solved by vanilla embeddings (e.g., tc07, tc12), there are gains of more than 10 percentage points in F1. It is further remarkable that on this problem, compared to link prediction and entity clustering, TuckER can also benefit from the protographs.

When looking at which protograph construction strategy is more suitable for which kind of problem, we see that for those test cases which use a concrete class in their definition (particularly tc07, tc08, tc11, and tc12), P2 is clearly superior to P1. This may be due to the fact that P1 does not create class embeddings for all classes used in the respective constructors. Therefore, the resulting entity embeddings are less discriminating.

5 CONCLUSION AND FUTURE WORK

In this paper, we present MASCHInE – a novel model-agnostic and protograph-assisted approach for learning KGEs. MASCHInE first trains embeddings on protographs and then transfers these embeddings to learn the final KGEs. We have shown that this approach produces more versatile KGEs, which can yield significant improvements in entity clustering and node classification. This is particularly noticeable for entity clustering, as protographs encode information about classes, on which entity clustering heavily relies on. We also demonstrated the advantage of using MASCHInE for link prediction: the use of protographs in the training procedure leads KGEMs to make more semantically valid predictions, which clearly highlights that geometric distance in the embedding space can be influenced by the semantics of the underlying KG.

In future work, we will extend MASCHInE by adding new protograph design principles. While in the current form, the transfer from the RDFS-based protograph embeddings to the KG entity embeddings is done once, we will study more expressive ontological constructs (e.g., OWL) and alternatives for KG and protograph iterative co-training. Additionally, it would be worth exploring whether pre-training over protographs provides benefits compared to passing the relevant RDFS triples to the KGEM as part of the whole graph. Comparing our approach with KGEMs that take advantage of additional information in the KG – e.g. attributes – also deserves further exploration.

REFERENCES

- [1] Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 5184–5193.
- [2] Peru Bhardwaj, John D. Kelleher, Luca Costabello, and Declan O’Sullivan. 2021. Poisoning Knowledge Graph Embeddings via Relation Inference Patterns. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 1875–1888.
- [3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Conf. on Neural Information Processing Systems (NeurIPS)*. 2787–2795.
- [4] Sejla Cebiric, François Goasdoué, Haridimos Kondylakis, Dimitris Kotzinos, Ioana Manolescu, Georgia Troullinou, and Mussab Zneika. 2019. Summarizing semantic graphs: a survey. *VLDB J.* 28, 3 (2019), 295–327.
- [5] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. 2018. HARP: Hierarchical Representation Learning for Networks. In *Proc. of the 32nd AAAI Conf. on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2127–2134.
- [6] Zijun Cui, Pavan Kapanipathi, Kartik Talamadupula, Tian Gao, and Qiang Ji. 2021. Type-augmented Relation Prediction in Knowledge Graphs. In *35th Conf. on Artificial Intelligence, AAAI 2021, 33rd Conf. on Innovative Applications of Artificial Intelligence, IAAI 2021, The 11th Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, February 2-9, 2021*. AAAI Press, 7151–7159.
- [7] Claudia d’Amato, Nicola Flavio Quatraro, and Nicola Fanizzi. 2021. Injecting Background Knowledge into Embedding Models for Predictive Tasks on Knowledge Graphs. In *The Semantic Web - 18th International Conf., ESWC 2021, June 6-10, Proc. (Lecture Notes in Computer Science, Vol. 12731)*. Springer, 441–457.
- [8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2D Knowledge Graph Embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 1811–1818.
- [9] François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. 2020. RDF graph summarization for first-sight structure discovery. *VLDB J.* 29, 5 (2020), 1191–1218.
- [10] Zengfeng Huang, Shengzhong Zhang, Chong Xi, Tang Liu, and Min Zhou. 2021. Scaling Up Graph Neural Networks Via Graph Coarsening. In *KDD ’21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*. ACM, 675–684.
- [11] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. 2022. Knowledge Graph Embeddings for Link Prediction: Beware of Semantics!. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG 2022) co-located with the 21th International Semantic Web Conference (ISWC 2022)*.
- [12] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. 2023. Sem@K: Is my knowledge graph embedding model semantic-aware?
- [13] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. 2023. Treat Different Negatives Differently: Enriching Loss Functions with Domain and Range Constraints for Link Prediction. *CoRR abs/2303.00286* (2023).
- [14] Nitisha Jain, Jan-Christoph Kalo, Wolf-Tilo Balke, and Ralf Krestel. 2021. Do Embeddings Actually Capture Knowledge Graph Semantics?. In *The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12731)*. Springer, 143–159.
- [15] Nitisha Jain, Trung-Kien Tran, Mohamed H. Gad-Elrab, and Daria Stepanova. 2021. Improving Knowledge Graph Embeddings with Ontological Reasoning. In *The Semantic Web - International Semantic Web Conf. ISWC, Vol. 12922*. 410–426.
- [16] Denis Krompaß, Stephan Baier, and Volker Tresp. 2015. Type-Constrained Representation Learning in Knowledge Graphs. In *The Semantic Web - 14th International Semantic Web Conf. (ISWC), Vol. 9366*. Springer, 640–655.
- [17] Maria Angela Pellegrino, Abdulrahman Altalba, Martina Garofalo, Petar Ristoski, and Michael Cochez. 2020. GEval: A Modular and Extensible Evaluation Framework for Graph Embedding Techniques. In *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12123)*. Springer, 565–582.
- [18] Marcin Pietrasik and Marek Z. Reformat. 2023. Probabilistic Coarsening for Knowledge Graph Embeddings. *Axioms* 12, 3 (2023), 275.
- [19] Jan Portisch and Heiko Paulheim. 2022. The DLCC Node Classification Benchmark for Analyzing Knowledge Graph Embeddings. In *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022 (Lecture Notes in Computer Science, Vol. 13489)*. Springer, 592–609.
- [20] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proc. of the 33rd International Conf. on Machine Learning, ICM, Vol. 48*. 2071–2080.
- [21] Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A Survey on Knowledge Graph Embeddings for Link Prediction. *Symmetry* 13, 3 (2021), 485.

- [22] Xueliang Wang, Jiajun Chen, Feng Wu, and Jie Wang. 2022. Exploiting Global Semantic Similarities in Knowledge Graphs by Relational Prototype Entities. *CoRR* abs/2206.08021 (2022). arXiv:2206.08021
- [23] Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation Learning of Knowledge Graphs with Hierarchical Types. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. IJCAI/AAAI Press, 2965–2971.
- [24] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR*.
- [25] Amal Zouaq and Félix Martel. 2020. What is the schema of your knowledge graph?: leveraging knowledge graph embeddings and clustering for expressive taxonomy learning. In *Proceedings of The International Workshop on Semantic Big Data, SBD@SIGMOD 2020, Portland, Oregon, USA, June 19, 2020*. ACM, 6:1–6:6.