Open-World Taxonomy and Knowledge Graph Co-Learning

Jiaying Lu (jiaying.lu@emory.edu), Carl Yang (j.carlyang@emory.edu)

Emory University, Atlanta, GA 30322, USA

Problem Formulation

TaxoKG Completion: Given an incomplete open-world TaxoKG, the task aims at inferring missing knowledge that is either entity-concept pair or entity-relation-entity triplet.

Motivations
1) TaxoKG contains taxonomic knowledge (AutoTaxo) and non-taxonomic knowledge (OpenKG);
2) The mutual enhancement between taxonomy and KG is helpful for TaxoKG completion;
3) TaxoKG is an open-world knowledge base that can easily accommodate fast-expanding data.

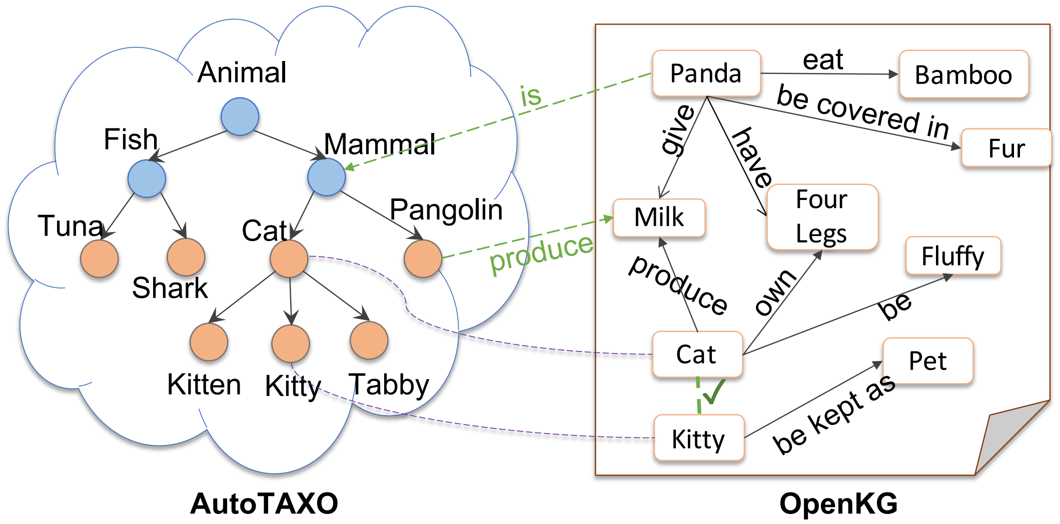
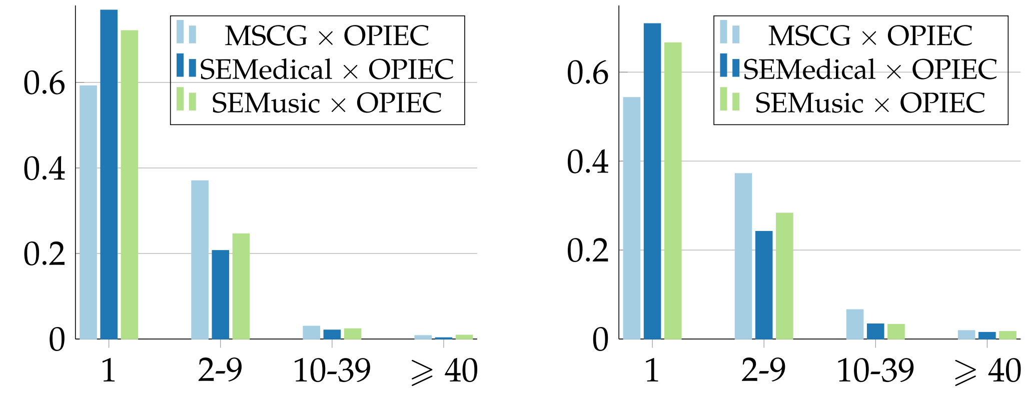


Figure 1: Toy example of TaxoKG.

Data Contribution

TaxoKG-Bench: A New Benchmark with Six Datasets for TaxoKG. TaxoKG-Bench is created from three AutoTaxos and two OpenKGs. We align AutoTaxos and OpenKGs by matching entities in entity-concept pairs with entities in entity-relation-entity triplets. Thus, TaxoKG-Bench contains many long-tailed entities and relations.



(a) Entities on * × OPIEC (b) Relations on * × OPIEC

Figure 2: Entity and relation histograms on three aligned TaxoKGs.

Benchmark Overview: TaxoKG-Bench is a large-scale, diverse, challenging benchmark. Our TaxoKGs typically contains thousands of relations, a great proportion of taxonomic knowledge, and significant amounts of unseen entities, concepts and relations.

Technical Contributions

We propose a novel model with the learn-to-conceptualize and learn-to-generalize abilities via combining **H**ierarchy-**A**ware **K**nowledge base **E**mbedding and **G**raph **C**onvolutional neural **N**etworks, namely HakeGCN. HakeGCN includes a series of essential technical designs for TaxoKG completion:

- The polar coordinates-based GCN encoder;
- The taxonomy-based sampling strategy;
- The GCN-oriented phased bounded decoder.

Overall Architecture of HakeGCN:

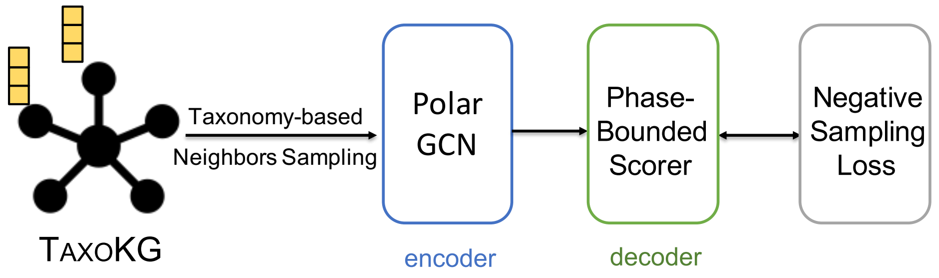


Figure 3: The unseen entities, concepts, and relations are handled by creating their embeddings from tokens of surface mentions. In the polar GCN encoder, both vertex and edge embeddings are first updated in Cartesian Coordinate via relational-GCN message propagation paradigm. Then, these updated embeddings are mapped into Polar Coordinate, which utilizes the modulus dimensions to reflect the depth of the taxonomy hierarchy and the phase dimensions to represent non-taxonomic relations. The phased-bounded decoder produces the validity score of input entity-concept pairs or entity-relation-entity triplets. To train HakeGCN, we collect positive knowledge examples and sample negative ones by proposed taxonomy-based neighborhood sampling.

Implementations: All code and data are publicly available at `https://github.com/lujiaying/Open-World-TaxoKG-CoLearning`.

Pilot Study on Combing Taxonomy and KG

Table 1: TaxoKG completion performance when presented with the separated data (SEMedical only or OPIEC only) v.s. the jointed data (SEMedical × OPIEC). Results validate the benefit of jointly modeling existing AutoTaxos and OpenKGs.

(a) Concept prediction results.

Model	Data	C-MAP	C-P@10, 30, 50
HAKE	AutoTaxo	.186	.344, **.355**, .177
HAKE	TaxoKG	**.262**	**.371**, .309, **.256**†
CompGCN	AutoTaxo	**.075**	**.284**, **.117**, **.109**
CompGCN	TaxoKG	.041	.060, .044, .032
HakeGCN	AutoTaxo	.105	.093, .093, .123
HakeGCN	TaxoKG	**.271**†	**.377**†, **.366**†, **.251**

(b) Relation prediction results.

Model	Data	R-MRR	R-H@10, 30, 50
HAKE	OKG	.350	**.454**, **.517**, **.545**
HAKE	TaxoKG	**.352**	.450, .509, .544
CompGCN	OKG	.006	.012, **.030**, **.049**
CompGCN	TaxoKG	**.009**	**.013**, .023, .034
HakeGCN	OKG	.375	.478, .555, .607
HakeGCN	TaxoKG	**.412**†	**.508**†, **.600**†, **.652**†

Experimental Evaluations

Baselines: We carefully adapt Translation-based methods (*i.e.* TransE, HAKE), Semantic matching-based methods (*i.e.* DistMult, HolE), GCN-based methods (*i.e.* R-GCN, CompGCN) , and Mutual enhancement-based method (*i.e.* LtCaG).

Protocols: There exist two sub-tasks for TaxoKG completion problem: AutoTaxo concept prediction and OpenKG relation prediction. For AutoTaxo concept prediction, we choose *Mean Average Precision* (MAP) and *Precision at N* (P@N) as evaluation metrics. For OpenKG relation prediction, we choose *Mean Reciprocal Rank* (MRR) and *Hits at N* (H@N) as metrics.

Table 2: TaxoKG completion results in different domains.

(a) General domain.

	MSCG × OPIEC			
	C-MAP	C-P@1, 3, 10	R-MRR	R-H@10, 30, 50
TransE	.006	.004, .002, .001	.002	.001, .004, .008
HAKE	.031	.014, .011, .010	.539	**.787**, **.821**, **.837**
DistMult	.001	9e-4, 3e-4, 3e-4	.080	.131, .159, .176
HolE	.006	.004, .002, .001	.002	.001, .004, .008
R-GCN	.044	.044, .017, .006	.017	.031, .121, .179
CompGCN	.004	.003, .002, .001	.011	.025, .051, .067
LtCaG	.003	.002, .001, .001	.002	.002, .006, .009
HakeGCN	**.070**	**.052**, **.027**, **.014**	**.675**	.756, .805, .832

(b) Music domain.

	SEMusic × ReVerb			
	C-MAP	C-P@1, 3, 10	R-MRR	R-H@10, 30, 50
TransE	.012	.053, .035, .028	.002	.002, .006, .009
HAKE	.201	.275, .270, .210	.131	.258, .344, .382
DistMult	.035	.118, .092, .066	.019	.039, .123, .188
HolE	.038	.118, .092, .066	.002	.002, .004, .007
R-GCN	.005	.011, .010, .013	8e-4	7e-4, .002, .003
CompGCN	.063	.092, .111, .095	.009	.019, .034, .042
LtCaG	182	.286, .251, .172	.003	.004, .006, .009
HakeGCN	**.238**	**.301**, **.307**, **.221**	**.178**	**.286**, **.412**, **.481**

More results: For more experimental results on remaining TaxoKGs, in-depth analysis, ablation study, and efficiency evaluations, please refer to our paper.

Table 3: Case Studies of Neighbors for Predicting Concepts and Relations.

(a) KG neighbors used in taxonomy concept prediction.

Concept	KG Neighbors
technique	(make from, recycled material, -) ✓ (architecture, be a thing of, -) ✓ (-, be apply, biology) ✓ (-, mean of, expression) ✗

(b) Taxonomy neighbors used in KG relation prediction.

Relation	Taxonomy Neighbors
be marry to	control ✗, family name ✓, guest ?
die from	illness ✓, disease ✓, disorder ✓
listen to	work of art ?, musical work ✓, piece of music ✓