Learning Class Definitions Using Ontology Embeddings*

First Author[1][0000−1111−2222−3333], Second Author[2,3][1111−2222−3333−4444], and Third Author[3][2222−−3333−4444−5555]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. The abstract should briefly summarize the contents of the paper in 15–250 words. Test citation [**Berners-Lee2001**]

Keywords: Ontology Embeddings · Class Definition · DBpedia.

1 Research Questions

This technical report aims to investigate whether knowledge graph embeddings are helpful in entity class definition in DBpedia by developing an experimental set-up. To answer this question, we calculate the link prediction scores for entities in pre-selected classes and compare them to the frequency-based baselines.

2 Empirical Semantics

The empirical semantics perspective of this work considers classes and properties of the DBpedia ontology. The main problematic aspect is the semantic expressivity and formalisation of the DBpedia categories (*skos:Concept*) derived from Wikipedia.

3 Introduction

[1 Page]
Explain your perspective on the problem of Empirical Semantics. Give both the intuition and motivate, by relying on use cases and examples, why this perspective is important. Briefly describe what is the state of the art and how you're pushing it with your contribution. Also mention what data and methods you use in your work. Conclude by clearly stating what is your contribution.

* Technical report of the task force 42 from ISWS 2022 led by Heiko Paulheim.

Problem definition by illustrative example.

Wikipedia categories define topics of resources in DBpedia. For example, the entity of type 'work' 'The Shining (novel)' [4] in DBpedia is defined through *dct:subject* property with several category URIs derived from Wikipedia, including the category 'Novels by Stephen King'[5]. This category, which is formally a *skos:Concept*, has such implicit information as type of written work, fiction genre ('Novel') and author ('Stephen King'). However, this implicit human-readable information included in a Wikipedia URI does not bring any additional formal semantics to the resource in DBpedia. Which properties of the resource 'The Shining (novel)' help automatically reason that it belongs to the category 'Novels by Stephen King'? What are the common properties of the resources, which are defined through the same Wikipedia category? Is it possible to detect such properties and assign new classes automatically, so that the nuanced categorisation of the resources can be made machine interpretable? This empirical semantics observations drive our research question and the experimental set-up, which aims to enrich the DBpedia ontology based on Wikidata categorisation.

About the method. Why do we focus on embeddings? We experiment with knowledge graph embeddings (in particular, RDF2vec) because..

About Related Work. There is work where embeddings were used to create class hierarchies, but not to predict classes (?) This is what novel in our direction.

About the data. For our experiment, we manually picked 13 categories from DBpedia that represent tangible and intangible real-world objects (such as books, persons, languages) and can be easily described with several properties. We collected all triples contained in the selected categories, taking only triples with DBpedia ontology predicates, since we aim at enriching this ontology. Additionally, we filtered out triples referring to Wikidata identifiers (such as 'dbo:wikiPageID') and triples with literal values as they are not relevant for our study. The detailed data overview is presented in Section 5.

Contributions. Results of the experiment.

4 Related Work

[1 Page]
List the main relevant work (a bullet list is ok) and for each of them write a paragraph describing (i) the key contribution of the related work, (ii) how your contribution relates/differentiate from it.

5 Resources

Our use-case focusses on DBpedia categories (why?). A category in DBpedia has a type of *skos:Concept*. There are more than 2 million categories in DBpedia. We manually selected 13 categories (motivation?) for our experiment. For every

[4] https://dbpedia.org/page/The_Shining_(novel)
[5] https://dbpedia.org/page/Category:Novels_by_Stephen_King

category, we manually selected predicate-object pairs that represent and distinguish a category. For example, the pair "author"-"Stephen_King" was chosen as an indicative of the category "Novels_by_Stephen_King". For six categories, more than one predicate-object pair was selected.

Four datasets were retrieved from DBpedia with SPARQL-queries:

1. Getting the number of subjects with the same predicate and object for every category. Criteria: (1) predicates are in the DBpedia ontology, (2) predicates not referring to Wikipedia (not containing the word "wiki" in their names). For example, there are 44 objects with predicate "author" and object "Stephen_King" in the category "Novels_by_Stephen_King". This predicate-object pair is the most frequent in the category and it is indicative for the category. The purpose of this dataset is to calculate the frequency of the predicate-object pairs in every category (ranking) and mean reciprocal rank (MRR). The overview of the selected categories, their indicative object-predicate pairs, ranks, number of predicate-object pairs, and the total number of triples is presented in Table 1.
2. Getting all triples for every category. Criteria: (1), (2), and (3) objects with URI containing "http://dbpedia.org/resource/" to filter out objects with literal values. This dataset is used for evaluation.
3. Getting all combinations of subjects and predicate-object-pairs in a category. Criteria: (1), (2), (3).
4. Counting the occurrences of the predicate-object pairs for every category in the whole DBpedia. This dataset is used in calculating TF-IDF.

Dataset 3: How many combinations we have for every category? See Table 2
Dataset 4: N of all the pairs in the whole DBpedia. See Table **??**

The resulting datasets, SPARQL-queries, and code are documented and openly available on GitHub[6].

Category	Predicate	Object	N Subjects	Rank	MRR	N P-O pairs	N Triples total
1990s_American_sitcoms	genre	Sitcom	273	1	1.000	3211	5865
Argentine_Primera_División_players	team	Argentina_national_football_team	660	5	0.200	6020	40958
Birds_of_Europe	class	Bird	1	1	1.000	12	12
	location	Europe	0	0			
English_pop_pianists	genre	Pop_music	7	1	1.000	227	255
	birthPlace	England	2	6	0.167		
	instrument	Piano	2	6	0.167		
Films_produced_by_Denzel_Washington	producer	Denzel_Washington	6	1	1.000	86	122
Hilltowns_in_Emilia-Romagna	region	Emilia-Romagna	2	1	1.000	6	8
	country	Italy	2	1	1.000		
	province	Province_of_Forlì-Cesena	1	3	0.333		
Italian_Renaissance_painters	movement	Italian_Renaissance	15	3	0.333	193	404
Languages_of_Namibia	spokenIn	Namibia	13	3	0.333	34	133
Novels_by_Stephen_King	author	Stephen_King	44	1	1.000	49	220
Philippine_television_talk_shows	genre	Talk_show	78	1	1.000	312	829
	country	Philippines	23	5	0.200		
Red_Hot_Chili_Peppers_songs	artist	Red_Hot_Chili_Peppers	49	1	1.000	209	643
Scottish_clans	country	Scotland	1	1	1.000	7	7
	type	Clan	0	0			
Swedish_death_metal_musical_groups	genre	Death_metal	61	1	1.000	506	960
	hometown	Sweden	24	3	0.333		

Table 1. DBpedia categories selected for experiment

[6] https://github.com/AndreiNesterov/42

Category	Predicate	Object	N Combinations
1990s_American_sitcoms	genre	Sitcom	
Argentine_Primera_División_players	team	Argentina_national_football_team	
Birds_of_Europe	class	Bird	
	location	Europe	
English_pop_pianists	genre	Pop_music	
	birthPlace	England	
	instrument	Piano	
Films_produced_by_Denzel_Washington	producer	Denzel_Washington	
Hilltowns_in_Emilia-Romagna	region	Emilia-Romagna	
	country	Italy	
	province	Province_of_Forlì-Cesena	
Italian_Renaissance_painters	movement	Italian_Renaissance	
Languages_of_Namibia	spokenIn	Namibia	
Novels_by_Stephen_King	author	Stephen_King	
Philippine_television_talk_shows	genre	Talk_show	
	country	Philippines	
Red_Hot_Chili_Peppers_songs	artist	Red_Hot_Chili_Peppers	
Scottish_clans	country	Scotland	
	type	Clan	
Swedish_death_metal_musical_groups	genre	Death_metal	
	hometown	Sweden	

Table 2. N of combinations of subjects and predicate-object pairs for every category

6 Proposed approach

[2 pages]
Describe your proposed method.

7 Evaluation and Results: Use case/Proof of concept - Experiments

[2 pages]
Show here that your proposed approach addresses your research questions or how you intend to show it. This can be done by either or both:

- Describing an experimental setting design, including research hypotheses, methods and metrics of measurements
- Describing a proof of concept/use case, based on real data, that support your claim

8 Discussion and Conclusions

[1 page]
Identify strengths and weaknesses of your proposal, discuss lessons learned: what are the key issues you have encountered or that you think should be taken into account to develop your proposal/experiments, and what are possible ways to address them.