Identifying global representative classes of DBpedia Ontology through multilingual analysis: A rank aggregation approach

Eun-kyung Kim and Key-Sun Choi

School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
{kekeeo,kschoi}@world.kaist.ac.kr

Abstract. Identifying the global representative parts from the multilingual pivotal ontology is important for integrating local language resources into Linked Data. We present a novel method of identifying global representative classes of DBpedia ontology based on the collective popularity, calculated by the aggregation of ranking orders from Wikipedia's local language editions.

Keywords: DBpedia, Multilingual, Ontology, Rank Aggregation

1 Introduction

The diversity and amount of data on the Web are both continuously growing, and there has been a paradigm shift leading from the publishing of isolated data to the publishing of interlinked data through a variety of knowledge sources such as Linked Open Data (LOD) [1]. DBpedia dataset [2] currently plays a central role in the LOD cloud, which has been populated using a large amount of collaboratively edited material (i.e., Wikipedia) as a knowledge source. Because of the ever-growing size and enormous scope of Wikipedia's coverage, the DBpedia dataset has been increasingly applied to a wide range of web applications.

The DBpedia dataset contains a community-curated cross-domain ontology to homogenize the description of information in the knowledge base (KB), which is one of the largest multilingual ontologies developed to date. Version 2014 of this ontology covers 685 classes in total, which form a subsumption hierarchy, and includes 2,795 different properties. This ontology has become a *de facto* reference vocabulary; however, this is limited as a multilingual pivot. Although a large number of instances among different languages are connected to the `owl:sameAs`[1] link, matching the class level is rare. The `rdfs:label` properties use language tagging to enhance multilingualism as follows.

[1] https://www.w3.org/TR/owl-ref/#sameAs-def

```
<owl:Class rdf:about="http://dbpedia.org/ontology/Actor">
    <rdfs:label xml:lang="en">actor</rdfs:label>
    <rdfs:label xml:lang="fr">acteur</rdfs:label>
    <rdfs:label xml:lang="ja">俳優</rdfs:label>
    <rdfs:label xml:lang="ko">영화인</rdfs:label>
    ...
```

This shows that the class "Actor" has several cross-lingual corresponding terms such as "영화인" in Korean and "Acteur" in French. Figure 1 shows the statistics of class numbers with `rdfs:label` properties. The number of labeled classes for different languages varies significantly, and there is obviously an absence of cross-lingual labeling for some editions such as Chinese. The DBpedia ontology (DBO) is continuously evolving due to its collaborative (wiki) paradigm and ongoing internationalization [3, 4]. However, it suffers from a scarcity of multilingual labels, due to its derivation that is based on the popular infoboxes in English. This leads to a limitation of other languages' ability to adapt the DBO to local language knowledge resources and makes it difficult to homogenize as a conceptual extension. Thus, identifying the global representative parts of the DBO is important for expanding multilingual ontologized space in LOD.

Figure 2 gives an overview of our motivation. Generally, the terminological components (henceforth referred to as the TBox) of an existent ontology can be translated and tailored to fit the understanding of other languages to expand multilingual coverage and thus increase knowledge access across languages with

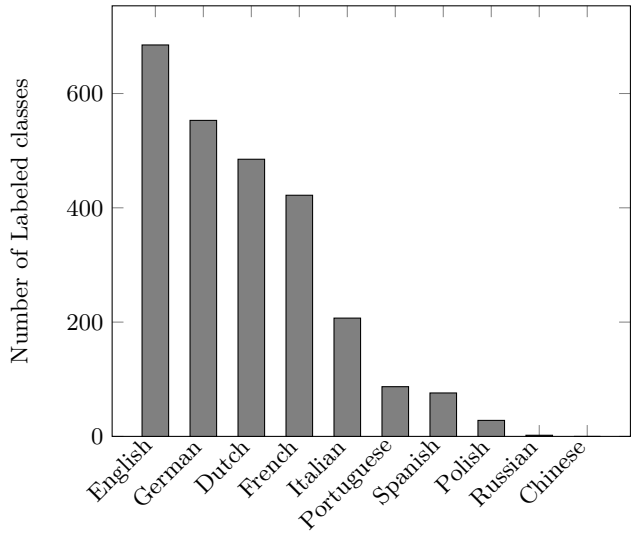

Fig. 1. Statistics for language-labeled classes of DBpedia ontology among 10 major languages

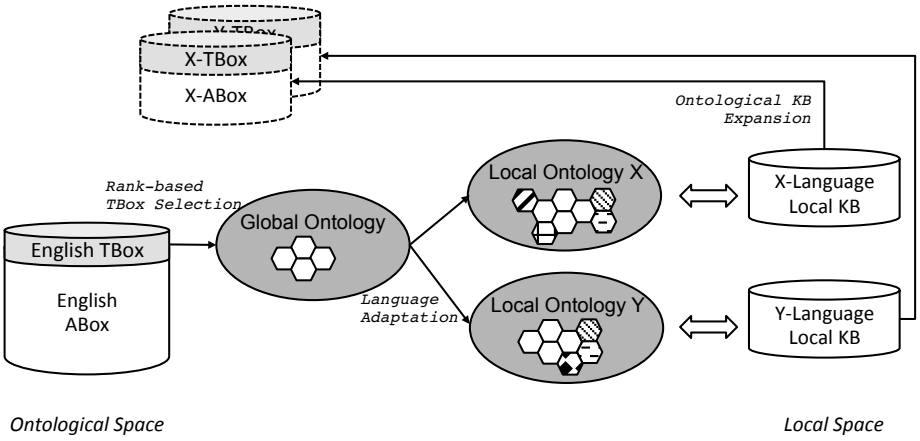

Fig. 2. Multilingual ontologized space expansion

existing ontologies [5]. Therefore, a multilingual pivotal ontology must accurately represent the global common concept structure, yet remain reusable in different languages so that connections can be made between local language knowledge resources and ontological KBs when entering an LOD.

We aimed to identify globally representative DBO classes for different language editions in this work, based on the combination of several ranking results that analyze the knowledge graph to measure the popularity of instances from multiple perspectives. Then, a consensus global ranking could be produced via rank aggregation; finally, we constructed a representative subset of DBO that could capture universally popular information that would be useful for improving the multilingual reuse of the ontology itself and would more easily and rapidly expand the ontological domain of the local language knowledge sources. We evaluated our approach by comparing its coverage with respect to the losses caused by the selection process, which had almost the same coverage with no appreciable loss of efficiency for larger sizes when the data were adapted to multilingual purposes.

2 Rank Aggregation–based Class Selection

When determining globally representative classes of DBO, we believe that the main challenge lies in the ranking model. Figure 3 shows an overview of the proposed approach that is mainly structured as two phases, as in the following subsections.

2.1 Language-Specific Popularity Analysis

We create a ranking model of classes to ascertain their degrees of significance in the ontology by analyzing each language dataset individually. We first computed

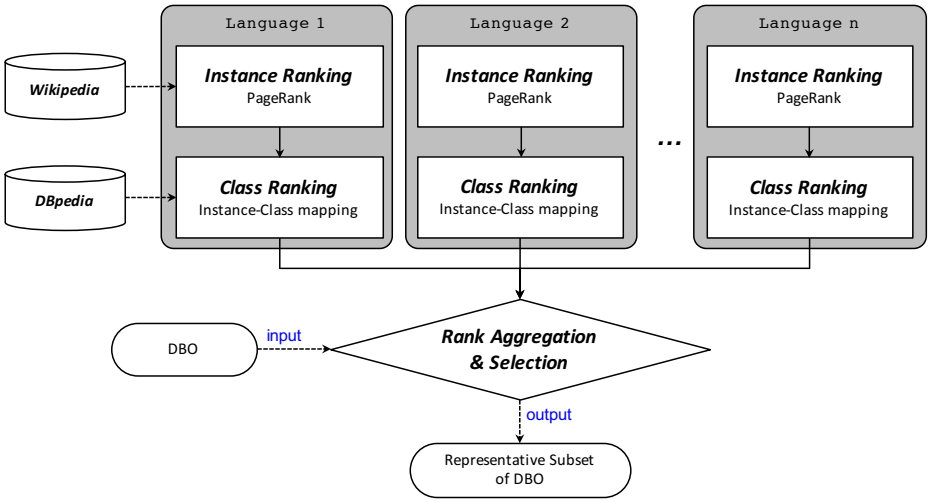

Fig. 3. Overview of the proposed framework

the ranking order for its instances and then combined these to determine the rank of a class. The rank of a class for each specific language is obtained by the PageRank [6, 7] values of its instances. We constructed a graph of instances from Wikipedia consisting of the links between articles to calculate the ranks of the instances. Each article corresponds to a node of the graph, and links between articles correspond to the edges of the graph.

Then, we calculated the rank of a class by mapping information between the instances in Wikipedia and the classes in DBpedia. We used "type" information from DBpedia to map instances to classes; for example, the instance "Barack Obama" is described and classified in three types as "OfficeHolder," "Person," and "Agent." Our class-level ranking model characterizes the following two features of a class to determine its rank:

1. A class is more popular if it is ranked higher based on the *average* of its instances' rank scores.
2. A class is more popular if it is widely *populated* in DBpedia ABox (i.e., the assertional component).

We used an aggregate function ($average * counting$) to compute the language specific class-level rank $CR^l(\mathtt{C})$ of a class $\mathtt{C}$ as:

$$CR^l(\mathtt{C}) = \frac{1}{n}\sum_{i=1}^{n} PR(i) * \sqrt{\|\mathtt{C}\|}, \tag{1}$$

where n is the number of instances of $\mathtt{C}$, $PR(i)$ is the PageRank score of instance i, and $\|\mathtt{C}\|$ indicates the unique number of populated instances of $\mathtt{C}$ in DBpedia ABox for a language l. In the results, $CR^l(\mathtt{C})$ represents the popularity of class $\mathtt{C}$ in the language l edition.

Table 1. Top 10 classes in different languages ranked by proposed approach; the distinct classes in each language are marked in bold type

English	French	Portuguese	Polish
Country	Country	Country	Country
Continent	**State**	Place	Place
Place	Continent	PopulatedPlace	PopulatedPlace
PopulatedPlace	**Department**	Agent	City
Agent	PopulatedPlace	Person	Agent
Organisation	Place	Organisation	Settlement
Person	Agent	Settlement	Person
Settlement	Person	City	Organisation
City	Settlement	**Artist**	**Region**
AmericanFootballTeam	**Territory**	**Work**	**AdministrativeRegion**

2.2 Language-Unified Popularity Analysis

The individual ranking orders from Section 2.1 are aggregated to produce a "globally popular" order of classes that would reflect their order of importance as judged by the collective evidence of all language editions. Table 1 depicts the top 10 independently ranked classes in four different language editions (the four sample languages in Figure 1). This means that different language editions of DBpedia may have different perspectives on the information that they contain. We produced the consensus rank for each language-specific ranking order using the existing score-based rank aggregation method (i.e., the Borda count method [8]). The Borda count is one of the most well-known and intuitive rank aggregation schemes in which each element for each ranking order is given a score depending on its rank, and these weights are then summed across all such ranking orders.

Each language-specific ranking is associated with a finite set of m classes $C = \{\mathtt{C}_1, ..., \mathtt{C}_m\}$, each of which is given a score depending on its place in the individual ranking order, the Borda scores are summed for all such individual scores to compute their total score. More formally, each class $\mathtt{C}_i$ has a different ranked position x, which is based on the class ranking function $CR^l(\mathtt{C}_i)$. We then define $\tau^{l_j}(\mathtt{C}_i) = x$ $(1 \leq x \leq m, 1 \leq j \leq n)$ such that the jth language edition ranks the class $\mathtt{C}_i$ at the xth position. Each class $\mathtt{C}_i$ has a Borda-based global ranking $CR^g(\mathtt{C}_i)$, defined as:

$$CR^g(\mathtt{C}_i) = \sum_{j=1}^{n}(m - \tau^{l_j}(\mathtt{C}_i)). \tag{2}$$

TBox Selection: After computing each class's global rank, the classes with the higher order global ranking scores are selected as the representative $\mathbb{R}$ with a certain size ρ by the following to define the classes and properties that should be included in:

Definition 1 *The set of classes* $C(\mathbb{R})$ *contains a class* $\mathtt{C}$ *iff:*

– $\mathtt{C}$ *is a class and* $CR^g(\mathtt{C}) \geq CR^g(\mathtt{C}_\rho)$ *or*
– $\mathtt{C}$ *is a class and there is a class* $\mathtt{D} \in C(\mathbb{R})$ *such that* $\mathtt{D} \sqsubseteq \mathtt{C}$

Definition 2 *The set of properties* $P(\mathbb{R})$ *contains a property* $\mathtt{p}$ *iff* $\mathtt{p}$ *is a property belonging to a class* $\mathtt{C} \in C(\mathbb{R})$.

3 Experimental Analysis

We measured the coverage of the representative subset; good representatives are expected to capture most of the information in the initial dataset without much loss. We compared the performance of our algorithm with two others: Monolingual-Rank (Mono) and Random-Selection (Rand). Mono is an approach that uses only English to calculate the rank computation. Rand represents the average performance of 10 runs by randomly selecting a subset of the dataset as a representative.

We used the DBpedia Mapping-based Dataset (2014) in our evaluation, which is a set of assertion triples that contain very specific information about the entities that can be used to query Wikipedia. Every instance in those triples is classified by the classes of DBO, and all properties are defined in the ontology. A sample RDF statement (in triple form: $< s, p, o >$) of this dataset that pertains to "Barack Obama" is as follows.

```
PREFIX dbo: http://dbpedia.org/ontology/
PREFIX dbr: http://dbpedia.org/resource/

<dbr:Barack_Obama, dbo:birthPlace, dbr:Hawaii>
```

This shows that the resource "Barack Obama" is the subject of other statements and presents a triple describing "Barack Obama"s birthplace as Hawaii.

We vary the number of representative subsets from 1 to 562 (the number of all classes involved in the rank; nearly 18% of the DBO's classes are never used in any languages) and compare the coverage achieved by the three methods listed in Table 2. It is clear from the results that the extracted globally popular classes have helped realize higher coverage for many languages.

3.1 Evaluations

We used gold standards[2] that were derived by assuming possible characteristics from both the number of the existing `rdfs:label` and the persistence of the classes. Then, we extracted a subset of DBO that could be adapted as the groundwork for automatic DBpedia mapping among languages through experimental evaluation.

[2] Evaluation data for this work is available for download at http://semanticweb.kaist.ac.kr/home/index.php/DBBO

Table 2. Coverage of representative set for ten languages defined in Figure 1. Percentages of triples are defined by classes in Ours, Mono, and Rand. $|\mathbb{R}|$ represents the size of the selected classes

$\|\mathbb{R}\|$	**Ours**	**Mono**	**Rand**
1	27.85%	0.85%	0.03%
5	67.54%	39.69%	0.03%
10	80.90%	67.54%	0.03%
20	84.54%	83.99%	2.26%
50	91.39%	84.60%	6.57%
100	93.05%	91.61%	12.96%
200	94.27%	92.31%	27.51%
300	94.74%	94.56%	42.35%
400	96.86%	94.74%	57.98%
500	97.13%	-	72.30%
562	97.18%	-	83.02%

For the first evaluation, we assumed that the classes that already have cross-lingual labels in many languages are important, because the labeling effort indicates their potential reuse in other languages. We create a gold standard based on this assumption by calculating "the number of existing `rdfs:labels` (Gold-standard 1)" for each class of DBO and divided the classes into two sets, positive and negative, in accordance with this assumption. 1) Positive set: Classes containing five (the mean value of DBO) or more labels. 2) Negative set: Classes containing fewer than five labels.

For the other evaluation, we assumed that the classes that are preserved across the two versions (the first and latest DBOs) are more important than the newly added classes in the 2014 version called "persistence (Gold-standard 2)." We divided the classes into the following two sets. 1) Positive set: Classes in both DBO 3.2 and DBO 2014. 2) Negative set: Classes that are only in DBO 2014. Figure 4 shows the F1 scores with respect to Gold-standard 1 and Gold-standard 2 as binary classifications. Based on these results, the size (ρ) of the multilingual pivotal ontology is set to 260 to obtain the best F1 score performance from the two evaluations. It is possible to reduce the size and hierarchy to only 260 from the top of the final order as a basis for the total of 685 classes. In comparison with Table 2, this smaller ontology may have at least approximately 90% coverage of ABox.

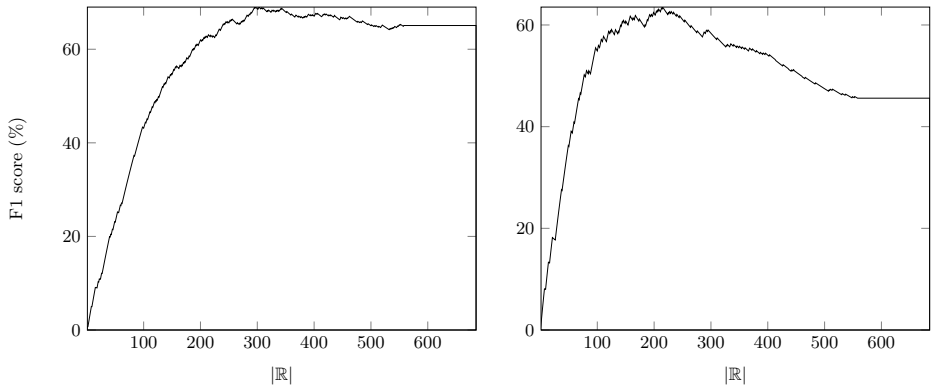

Fig. 4. F1 scores of Gold-standard 1 (the left side) and Gold-standard 2 (the right side) with respect to $|\mathbb{R}|$

4 Conclusion

We presented an approach for identifying global representative classes from DBpedia ontology (DBO), regarded as a multilingual pivotal ontology in this work. We combined the different independently constructed preferences of ranks for each language edition of Wikipedia to produce a consensus order of classes for DBO that is more desirable for representing the knowledge base for multilingual reuse and for connectability as Linked Open Data (LOD) through ontology. Our experimental results showed that the proposed approach significantly helped improve the labeling performance for non-English languages compared to both monolingual and random methods; the selected classes can be smaller than the entire ontology without significant loss of coverage. We expect that a representative subset of DBO in this paper will have a central role in the enrichment and integration of local language knowledge resources in LOD, avoiding islands of monolingual data.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP)(No. R0101-16-0054,WiseKB: Big data based self-evolving knowledge base and reasoning platform); the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP(2015M3A9A7029735).

References

1. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

2. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.
3. Eun-Kyung Kim, Matthias Weidl, Key-Sun Choi, and Sören Auer. Towards a korean dbpedia and an approach for complementing the korean wikipedia based on dbpedia. In Sören Auer, Jonathan Gray, Claudia Müller-Birn, Rufus Pollock, and Sara Wingate Gray, editors, *Proceedings of the 5th Open Knowledge Conference*, volume 575, pages 12–21. CEUR-WS.org, 2010.
4. Dimitris Kontokostas, Charalampos Bratsas, Sören Auer, Sebastian Hellmann, Ioannis Antoniou, and George Metakides. Internationalization of linked data: The case of the greek dbpedia edition. *J. Web Sem.*, 15:51–61, 2012.
5. Jorge Garcia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11(0), 2011.
6. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1999.
7. Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.
8. J. C. Borda. Memoire sur les elections au scrutin, 1781.