Discovering Names in Linked Data Datasets

Bianca Pereira, João C. P. Da Silva, Adriana S. Vivacqua
Federal University of Rio de Janeiro

Outline

- Motivation
  - Tasks
  - How to find what I need?
- Our goal
- Approaches to find named entities and their names
  - Manual
  - Rdfs:label
  - Heuristics
- Evaluation of our heuristics
- Conclusion

GRECO
GRUPO DE ENGENHARIA DO CONHECIMENTO - PPGI/UFRJ

Named Entity Resolution / Entity Linking

- Find Named Entities and their names in a text.
- Find Named Entities and their names in a dataset.

- Match them.

Using Linked Data Datasets

- Named Entities are not identified naturally in the schema.
- Names are literals in a property's range.
- Properties may vary from dataset to dataset.

How to find named entities in a Linked Data Dataset

- Named Entities are identified by the same classes. (Person, Organization and so on)
- For each class there are properties that identify names.

Goal

- Find PIN (Properties that Identify Names) in generic Linked Data datasets.

Methods to find Named Entities

- Manual
- Using rdfs:label
- Using heuristics

Finding named entities manually

- First, it requires a lot of manual work.
- Linked Data datasets may have a huge schema.
- A class that identifies Named Entities in a dataset sometimes does not identify them in another dataset.
- PINs from a class change from a dataset to another.

Using rdfs:label

It is a natural choice because:

- It is part of RDF Schema.
- It refers to a “human-readable name” for RDF resources.

Rdfs:label in Linked Data datasets

Features	Linked MDB	Geo Linked Data	Linked Brainz	Jamendo
Use of rdfs:label.	almost all classes	always	few classes	never
Rdfs:label used for names.	partially	Always with restrictions	always	-
Are there other PINs?	yes	no	yes	yes

Using heuristics

- Sometimes datasets do not use “default PINs”.
- Sometimes datasets use different PINs for the same class.
  One dataset may use many vocabularies that are not directly related.
- Datasets use classes in a different way.
  mo:Signal is used by Linked Brainz and Jamendo. In the first it does not identify named entities. In the latter it does.

Using heuristics

- We developed 4 different heuristics to identify PIN for each class:
  - Naive
  - Parametrized Naive
  - Multivalue
  - Multivalue with Threshold

- Each class with at least one PIN can be considered as identifying Named Entities.

Using heuristics

- To calculate the score for each PIN we:
  1. get all classes
  2. for each class:
     2.1. get all properties
     2.2 for each property:
     2.2.1 analyze the value according to the heuristic.

- Values are considered if they are Proper Names. We are considering literals where at least 50% of words are started by a capitalized letter.

Using Heuristics

Naive	· Properties with the highest occurrence of Proper Names for each class are PINs
Parametrized Naive	- Properties with the highest occurrence of Proper Names with more than *min* characters and less than *max* characters for each class are PINs.
Multivalue	- Every property with at least one occurrence of Proper Names with less than *max* characters is considered a PIN.
Multivalue with Threshold	- Every property with at least one occurrence of Proper Names with less than *max* characters is considered a PIN if it appears in more than *threshold%* of instances of a class.

Experiments

- Gold Standard
  - Preferential names
  - Alternative names
  - Acronyms

- Tested with datasets
  - Jamendo (DBTune.org)
  - Linked Movie Database

Results

Method	Jamendo		Linked Movie Database	
	PIN found	False positives	PIN found	False positives
Naive	3 (100%)	1	31 (75.61%)	21
Parametrized Naive	3 (100%)	1	19 (46.34%)	33
Multivalue	3 (100%)	3	41 (100%)	85
Multivalue with Threshold (0.4)	2 (66.67%)	0	35 (85.36%)	68
Multivalue with Threshold (0.6)	2 (66.67%)	0	35 (85.36%)	65
Multivalue with Threshold (0.8)	2 (66.67%)	0	34 (82.93%)	61
Multivalue with Threshold (0.95)	0	0	32 (78.05%)	54

GRECO
GRUPO DE ENGENHARIA DO CONHECIMENTO - PPGI/UFRJ

Results

Heuristics	Jamendo			Liked Movie Database		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Naive	0.75	**1**	0.8571	**0.5962**	0.7561	**0.6667**
P. Naive	0.75	**1**	0.8571	0.3654	0.4634	0.4085
Multivalue	0.5	**1**	0.6667	0.3254	1	0.4910
Multivalue (0.4)	**1**	**1**	**1**	0.3398	**0.8536**	0.4861
Multivalue (0.6)	**1**	**1**	**1**	0.35	**0.8536**	0.4964
Multivalue (0.8)	**1**	**1**	**1**	0.3579	0.8293	0.5
Multivalue (0.95	0	0	0	0.3721	0.7805	0.5039

GRECO
GRUPO DE ENGENHARIA DO CONHECIMENTO - PPGI/UFRJ

Conclusion

- Rdfs:label is useful but insufficient to find Named Entities's name.
- The use of heuristics is a feasible and good way to find named entities and their PIN in a dataset with generic schema.
- Even a simple heuristic is capable of find PINs in a generic Linked Data dataset.

Thanks!

Bianca de Oliveira Pereira
bianca.pereira@ppgi.ufrj.br
bianca.oli.pereira@gmail.com
Skype: bianca.oli.pereira