Automatic Extraction of Axioms from Wikipedia Using SPARQL

Lara Haidar-Ahmad, Amal Zouaq, Michel Gagnon
Ecole Polytechnique de Montreal, Department of Computer and Sowtware Engineering,
University of Ottawa, School of Electrical Engineering and Computer Science
lara.haidar-ahmad@polymtl.ca , azouaq@uottawa.ca , michel.gagnon@polymtl.ca

Abstract. Building rich axiomatic ontologies automatically is a step towards the realization of the Semantic Web. In this paper, we describe an automatic approach to extract complex classes' axioms from Wikipedia definitions based on recurring syntactic structures. The objective is to enrich DBpedia concept descriptions with formal definitions. We leverage RDF to build a sentence representation and SPARQL to model patterns and their transformations, thus easing the querying of syntactic structures and the reusability of the extracted patterns. Our preliminary evaluation shows that we obtain satisfying results, which will be further improved.

1 Introduction

Building rich ontologies with reasoning capabilities is a difficult task, which can be time consuming. It requires both the knowledge of domain experts and the experience of ontology engineers. This is one of the main reasons why current Semantic Web and linked data rely mostly on lightweight ontologies. The automatization of axiom extraction is a step towards creating richer domain concept descriptions [4] and building a Semantic Web that goes beyond explicit knowledge for query answering. Ontology learning, i.e. the automatic extraction of ontologies from text, can help automatize the extraction of primitive, named and complex classes. Few state of the art approaches were developed to achieve this goal, mostly pattern-based approaches [2, 3]. To our knowledge, LExO [3] is the most advanced system for complex class extraction. This paper describes our approach to extract defined and primitive class axioms from Wikipedia concept definitions using SPARQL. The main contribution of this work is i) the utilization of SPARQL graph matching capabilities to model patterns for axiom extraction ii) the description of SPARQL patterns for complex class extractions from definitions and iii) The enrichment of DBpedia concept descriptions using OWL axioms and defined classes. We also briefly compare our preliminary results with those of LEXO.

2 Methodology

We rely on a pattern-based approach to detect syntactic constructs that denote complex class axioms. These axioms are extracted from Wikipedia definitions.

Definition Representation and General Pipeline: We process definition sentences and first construct an RDF graph that represents the dependency structure of the definition and the words' part of speech and positions in the sentence. This step makes the subsequent step of pattern matching using SPARQL requests easier. For every word, we specify its label, its part of speech, its position in the sentence and its grammatical relations with the other words based on the output of the Stanford parser [1]. Figure 1 presents an example of the RDF graph of a definition. For this example, we use the definition of the Wikipedia concept *Vehicle* from our dataset, which is "Vehicles are non-living means of transportation".

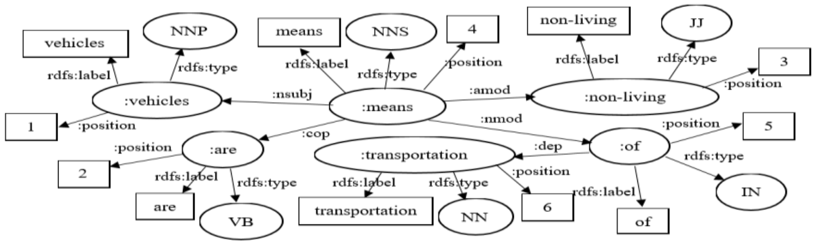

Fig. 1. The RDF Representation of the definition of *vehicles*.

Based on this RDF representation, we execute a pipeline of SPARQL requests on the obtained RDF graphs (see Figure 2).

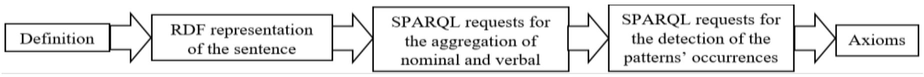
Definition
RDF representation of the sentence
SPARQL requests for the aggregation of nominal and verbal
SPARQL requests for the detection of the patterns' occurrences
Axioms

Fig. 2. Pipeline.

First, we execute SPARQL aggregation requests to extract complex expressions such as nominal and verbal groups and define subclass axioms. For instance, for the sentence *vehicles are non-living means of transportation*, we obtain the following expressions: *vehicles*, *non-living means of transportations* and *means of transportations*. We also extract the axiom *subClassOf* (*Non-living means of transportation, Means of transportation*). Finally, we execute a set of SPARQL axiom queries to identify occurrences of patterns that can be mapped to OWL complex class definitions.

SPARQL Pattern Representation: Based on a randomly chosen set of 110 definitions from Wikipedia and their sentence representation, we identified several recurring syntactic structures manually and built their corresponding SPARQL patterns. Next, we mapped patterns to complex class axioms using SPARQL

CONSTRUCT. Table 1 presents the most common patterns that we identified in our dataset, in addition to their corresponding axioms. Each pattern is modeled using a single SPARQL request. This mechanism provides simple ways to enrich our approach with patterns that we do not support yet.

Table 1. Most Frequent Patterns and their Respective Axioms.

Frequent patterns for the definitions of concepts	Corresponding Axioms
(1) SUBJ copula COMP *Vehicles are non-living means of transportation.*	$\text{SUBJ} \subseteq \text{COMP}$ $\textit{Vehicles} \subseteq \textit{NonLivingMeansOfTransportations}$
(2) SUBJ copula COMP that VERB OBJ *A number is an abstract entity that represents a count or measurement*	$\text{SUBJ} \equiv (\text{COMP} \cap \exists \text{VERB.OBJ})$ $\textit{Number} \equiv (\textit{AbstractEntity} \cap \exists \textit{represents.(Count} \cup \textit{Measurement))}$
(3) SUBJ copula COMP VERB preposition NOUN *A lake is a body of water surrounded by land.*	$\text{SUBJ} \equiv (\text{COMP} \cap \exists \text{VERB\_prep\_NOUN})$ $\textit{Lake} \equiv (\textit{BodyOfWater} \cap \exists \textit{surroundedBy.Land})$

3 Preliminary Evaluation and Discussion

We compared the generated axioms with a manually-built gold standard containing 20 definitions chosen randomly from our initial dataset[1]. We assessed the correctness of the axioms using standard precision and recall by focusing on named classes, predicates and complete axioms (see Table 2). Complete axioms metrics are calculated by counting the number of classes, predicates and logical operators matched with the ones in the gold standard. We obtain a macro precision and recall of 0.86/0.59 respectively. We also propose an axiom evaluation based on the Levenstein similarity metric which considers each axiom as a string. The higher the Levenstein similarity between the generated axiom and the reference, the most similar the axioms are. We tested multiple similarity levels as shown in Table 3. We notice that we usually generate the right axioms for i) small sentences ii) sentences with a simple grammatical structure and iii) longer sentences which have no grammatical ambiguities. We also notice that false positives are rarely generated, and the errors in our results are usually caused by incomplete axioms. This is explained by the limited number of implemented patterns (10 patterns).

Table 2. Evaluation Results.

	Classes		Predicates		Complete Axioms	
	Precision	Recall	Precision	Recall	Precision	Recall
Macro	0.87	0.66	0.94	0.54	0.86	0.59
Micro	0.86	0.61	0.76	0.36	0.78	0.48

[1] The dataset and gold standard are available at: http://westlab.herokuapp.com/axiomfactory/dataESWC16

Table 3. Axioms' Precision based on Levenstein Similarity with the Gold Standard.

Similarity Level	0.70	0.80	0.90	1.00
Levenstein Precision	0.55	0.50	0.35	0.30

While LExO [3] adopted a similar approach to ours, they did not rely on standard Semantic Web languages such as SPARQL for their patterns and did not take into account the aggregation of nominal and verbal groups, or the extraction of taxonomical relations. For example, given the definition *A minister or a secretary is a politician who holds significant public office in a national or regional government*, LExO generates (*Minister ∪ Secretary) ≡ (Politician ∩ ∃holds.((Office ∩ Significant ∩ Public) ∩ ∃in.(Government ∩ (National ∪ Regional))))*. In contrast, our system generates the axiom *Minister ≡ Secretary ≡ (Politician ∩ ∃ holds.(SignificantPublicOffice ∩ ∃in.(NationalGovernment ∪ RegionalGovernment)))*, and in addition, it generates a taxonomy where, *SignificantPublicOffice* is a subclass of *PublicOffice*, and *NationalGovernment* and *RegionalGovernment* are subclasses of *Government.*

4 Conclusion and Future Work

The paper describes an approach to extract OWL axioms with the aim to *logically define* DBpedia concepts from Wikipedia definitions using SPARQL requests. We are currently working on the implementation of our pipeline as a Web service, which has not been proposed yet in the state of the art. More importantly, one original contribution of this paper is the reliance on Semantic Web languages (RDF, SPARQL) to model sentences, patterns and axioms, thus easing the reusability and enrichment of the defined patterns.

Acknowledgement

This research has been funded by the NSERC Discovery Grant Program.

References

1. De Marneffe, M-C, and Manning, C. D. (2008). The Stanford typed dependencies representation. Proc. of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation. ACL.
2. Bühmann, L., Fleischhacker, D., Lehmann, J., Melo, A. & Völker, J. (2014). Inductive Lexical Learning of Class Expressions. In Knowledge Engineering and Knowledge Management (pp. 42-53). Springer International Publishing.
3. Völker, J., Haase, P. and Hitzler, P. (2008) Learning expressive ontologies, in: Proc. of the Conference on ontology Learning and Population, IOS Press, 2008, pp. 45–69
4. Font, L., Zouaq, A., Gagnon, M. (2015). Assessing the Quality of Domain Concepts Descriptions in DBpedia. SITIS 2015: 254-261