Verifying Integrity Constraints of a RDF-based WordNet

Fabricio Chalub

Alexandre Rademaker

IBM Research, Brazil

January 30, 2016

OpenWordnet-PT

http://wnpt.brlcloud.com/wn/

- Goal: not a simple translation of PWN, based on PWN architecture.
- originally created from a (PT) projection of the Universal WordNet (Gerard de Melo)
- Three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries.
  - Corpora: AC/DC project, DHBB CPDOC/FGV etc.
  - LR: morphosemantic links, nominalizations from NomLex, Nomage and Wiktionary etc.
- Freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
- used by "Google Translate", FreeLing, OMW, BabelNet and Onto.PT.

Why RDF

```
<definition gloss="This is my definition">
<meta creator="me/"></definition>

<definition><meta creator="me"/>
This is my definition</definition>

<definition><meta creator="me"/>
 <text>This is my definition</text>
</definition>

<lmf:definition meta:creator="me">
This is my definition</lmf:definition>
```

“Why RDF model is different from the XML model” by Tim Berners-Lee (1998). http://www.w3.org/DesignIssues/RDF-XML.html

Why RDF (cont.)

- There is a mapping from XML documents to semantic graphs.
- The element names were a big hint for a human reader.
- Without the schema (DTD, XML Schema), you know things about the doc structure, but nothing else. You can't tell what to deduce.
- You can't even really tell what real questions can be asked.
- (1) mapping is many to one; (2) you need a schema to know what the mapping is (don't have a inference language); (3) the expression you need for querying something in terms of the XML tree is necessarily more complicated than the expression you need for querying something in terms of the RDF tree.
- "give me the properties with the same metadatas of this one?"

Why RDF (cont.)

- Giving a machine a knowledge tree vs. giving a person a document.
- A document for a person is generally serialized so that, when read serially by a human being, the result will be to build up a graph of associations in that person's head. The order is important.
- For a graph of knowledge, order is not important, so long as the nodes in common between different statements are identified consistently.

Linked Data

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- Include links to other URIs. so that they can discover more things (ILI?)

A number of Linked Data projects for lexical resources.

“Linked Data” by Tim Berners-Lee (2006).

Some words about vocabularies

- To encode data, we need to decide which classes and properties to use!
- Different vocabularies for RDF encoding wordnets!
- The adoption of already defined vocabularies helps on the data interoperability since these makes data easily integrate with other resources.
- We use `http://www.w3.org/TR/wordnet-rdf/` from 2006.

Scripts available `http://github.com/own-pt/`.

Namespaces

```
https://w3id.org/own-pt/wn30/schema/
https://w3id.org/own-pt/wn30-pt/instances/
https://w3id.org/own-pt/wn30-en/instances/
https://w3id.org/own-pt/nomlex/schema/
https://w3id.org/own-pt/nomlex/instances/
```

This paper

Our first attempt at verifying integrity constraints of our openWordnet-PT against the ontology for Wordnets encoding.

Correcting and improving linguistic data is a hard task.

So far, no clear criteria for semantic evaluation wordnets not ways of comparing their relative quality or accuracy. Thus qualitative assessment of a new wordnet seems, presently, a matter of judgment and art.

OWL and RDF

- Consistency check of OWL and Integrity Constraints in RDF
- OWL Lite and OWL DL semantics are based on Description logics. DL are a family of logics that are decidable fragments of first-order logic with attractive and well-understood computational properties.
- A DL knowledge base is comprised by two components, TBox and ABox. The TBox contains intensional knowledge (terminology).
- The ABox contains extensional knowledge (assertional).
- Intensional knowledge is usually thought not to change and extensional knowledge is usually thought to be contingent.

Reasoning

Given an ontology encoded in OWL (Lite or DL) one can use DL reasoners for different tasks such as: concepts consistency checking, query answering, classification, etc.

The basic reasoning task in an ABox is instance checking, which verifies whether a given individ- ual is an instance of (or belongs to) a specified concept.

In some use cases, we want a method to validating the RDF data regarding a given model. In this case, OWL users intend OWL axioms to be interpreted as constraints on RDF data.

CWA vs. OWA

OWL default semantics adopts the Open World Assumption (OWA) and does not adopt the Unique Name Assumption (UNA).

Due to OWA, a statement must not be inferred to be false on the basis of failures to prove it; the fact that a piece of information has not been specified does not mean that such information does not exist.

On the other hand, the absence of UNA allows two different constants to refer to the same individual.

Queries in SPARQL

One more motivation for RDF:

```
 select ?w ?ws1 ?ws2
{
  ?ss1 wn30:containsWordSense ?ws1 .
  ?ws1 wn30:word ?w .
  ?ss2 wn30:containsWordSense ?ws2 .
  ?ws2 wn30:word ?w .
  ?ss1 wn30:hyponymOf* ?ss2 .
}
```

Tools

Protege is an ontology editor that among other features has interface with two well-know DL reasoners: FaCT++, HermiT etc.

Starting in version 4, Protege also gives us an interface to search for explanations that caused an inconsistency. Racer and Pellet are reasoners that have this feature builtin.

RDF_{pro} for combine, split and syntax check etc.

Stardog and Allegro Graph triplestores. Stardog has ICC included, AG has RDF++ semantics.

Errors

Errors found can be categorized in three different classes: datatype errors, domain and range errors, structural errors.

Missing classes and properties definitions. We improved the OWL file.

```
wn30:AdjectiveWordSense rdfs:subClassOf
 wn30:WordSense .
```

Literal values with types.

```
Literal value "00113726" does not
 belong to datatype nonNegativeInteger
```

Errors

current account is the label of wordsense-13363970-n-3 and Britain the label of wordsense-08860123-n-4.

```
Explanation for:
 Thing SubClassOf Nothing
  classifiedByRegion Domain Synset
  current_account classifiedByRegion Britain
  current_account Type WordSense
  Synset DisjointWith WordSense
```

```
wordsense-13363970-n-3 classifiedByRegion
   wordsense-08860123-n-4
```

> *"The following pointer types are usually used to indicate lexical relations: Antonym, Pertainym, Participle, Also See, Derivationally Related. The remaining pointer types are generally used to represent semantic relations."*

Fixing Errors

```
wn30:classifiedByRegion
 a rdf:Property, owl:ObjectProperty ;
 rdfs:domain wn30:Synset ;
 rdfs:range wn30:NounSynset ;
 rdfs:subPropertyOf wn30:classifiedBy .
```

Updated to:

```
wn30:classifiedByRegion
  a rdf:Property, owl:ObjectProperty ;
  rdfs:subPropertyOf wn30:classifiedBy ;
  rdfs:range [ a owl:Class ;
    owl:unionOf (wn30:NounWordSense
               wn30:NounSynset)] ;
  rdfs:domain [ a owl:Class ;
    owl:unionOf (wn30:WordSense wn30:Synset)] .
```

Proofs Explanations

In formal verifications, the complexity of the proofs/explanations. This is the explanation found for the issue:

```
synset-01345109-v hypernymOf
    synset-01220528-v
 VerbWordSense subClassOf WordSense
 frame domain VerbWordSense
 synset-01220528-v frame
    "Somebody ----s something"
 hypernymOf range Synset
 Synset disjointWith WordSense
```

`synset-01220528-v` found to be of type 'Synset' due to it is the object of a triple with predicate `wn30:hypernymOf` combined with the range of this predicate is the set of all synsets.

`synset-01220528-v` is a verb synset and that verb synsets are a subset of synsets.

More

Two invalid situations: (a) two or more words associated to a single word sense subject; (b) two or more lexical forms associated to a single word subject.

```
wordsense-01860795-v-2 type WordSense
word-deixar lexicalForm "deixar"@pt
word-parar lexicalForm "parar"@pt
wordsense-01860795-v-2 word word-deixar
Word subClassOf lexicalForm exactly 1
wordsense-01860795-v-2 word word-parar
word-deixar type Word
word-parar type Word
WordSense subClassOf word exactly 1 Word
```

More

Stardog is the only reasoner and database system that supports ICV. Under the ICV semantics, the axioms below from the *wn30:WordSense* class were taken as constraints rather than terminology definitions. Finding an instance of the class *wn30:WordSense* connected to more than one instance of *wn30:Word*, it will raise an exception instead of infer that the two different *wn30:Word* instances should be the same.

```
wn30:WordSense
  a rdfs:Class, owl:Class ;
  rdfs:subClassOf [
   a owl:Restriction ;
   owl:onProperty wn30:inSynset ;
   owl:qualifiedCardinality
     "1"^^xsd:nonNegativeInteger ;
    owl:onClass wn30:Synset ], [
   a owl:Restriction ;
   owl:onProperty wn30:word;
   owl:qualifiedCardinality
    "1"^^xsd:nonNegativeInteger ;
   owl:onClass wn30:Word ] .
```

Conclusions

- Linguistic resources are very easy to start, hard to improve and extremely difficult to maintain.
- Size of lexical resources are easy to compare, quality is hard.
- A lot of (old? already used/defined?) verifications can be encoded in OWL axioms. Some of them may require more expressivity (SUMO?)