Towards the Natural Ontology of Wikipedia

Andrea Giovanni Nuzzolese[1,2], Aldo Gangemi[1,3],
Valentina Presutti[1], and Paolo Ciancarini[1,2]

[1] STLab-ISTC, National Research Council, Rome, Italy.
[2] Dept. of Computer Science and Engineering, University of Bologna, Italy.
[3] LIPN, University Paris 13, Sorbone Cité, UMR CNRS, France

Abstract. In this paper we present preliminary results on the extraction of ORA: the Natural Ontology of Wikipedia. ORA[4] is obtained through an automatic process that analyses the natural language definitions of DBpedia entities provided by their Wikipedia pages. Hence, this ontology reflects the richness of terms used and agreed by the crowds, and can be updated periodically according to the evolution of Wikipedia.

1 An ontology for Wikipedia

The DBpedia Ontology [5] (DBPO) and Yago [7] are the two reference ontologies for DBpedia. Both of them provide only partial extensional and intensional coverage of DBpedia entities because they rely on Wikipedia categories (Yago) and infoboxes (DBpedia), which induce an intrinsic limit of domain coverage [3]. Tìpalo [3] is a tool that automatically produces a RDF taxonomy of types (aligned with WordNet and Dolce) for a DBpedia entity by analysing its natural language definition in Wikipedia. This approach is aimed at identifying the most natural types for an entity as they are expressed by the crowds. By running Tìpalo on the whole DBpedia we aimed at deriving a natural ontology for Wikipedia and approximating as much as possible a complete domain coverage. In this paper, we show the results obtained so far from this process: the first version of the Natural Ontology of Wikipedia (ORA)[6], and we discuss emerging issues and possible solutions for its refinement. This article is organized as follows: (i) in section 2, we introduce the main related work; (ii) in section 3 we describe the material and the method used for generating the ontology; (iii) finally in section 4 we describe the results obtained so far, and discuss ongoing work and future research directions.

2 Related work

The DBpedia project [4] and YAGO [7] are the most relevant approaches at generating an ontology from semi-structured information in Wikipedia. DBpe-

[4] ORA is the italian translation of NOW

[5] http://dbpedia.org/ontology

[6] http://isotta.cs.unibo.it:8080/sparql - select the graph *now*

dia provides an ontology extracted from Wikipedia infoboxes based on hand-generated mappings of infoboxes to the DBpedia ontology (DBPO). DBPO counts 359 concepts (version 3.8) but only 2.3M entities over more than 4M are classified with respect to this ontology. YAGO types are extracted from Wikipedia categories and aligned to a subset of WordNet. The YAGO ontology is larger that DBPO and counts $\sim$290K concepts. YAGO has a larger (although still incomplete, 2.7M typed entities) coverage of DBpedia entities. ORA introduces a third *dimension*: the terminology of the crowds; furthermore, it provides a larger coverage (currently 3.0M typed entities). Recently, the Schema.org [7] initiative has provided alignments to the DBPO. However, such effort does not add value from the perspective of the intensional and extensional coverage issues. Other relevant work related to our method includes Ontology Learning and Population (OL&P) techniques [1]. Typically OL&P is implemented on top of machine learning methods, hence it requires large corpora, sometimes manually annotated, in order to induce a set of probabilistic rules. Such rules are defined through a training phase that can take a long time. The method used for ORA and implemented by Tìpalo [3] differs from existing approaches as it is mainly rule-based, hence it does not require a training phase and it is faster than the other approaches.

3 Automatic extraction of an ontology for Wikipedia: materials and methods

Tìpalo is implemented as a pipeline of components and data sources. Each component in the pipeline implements a step of the computation: (i) extraction of an entity's natural language definition from its Wikipedia abstract; (ii) natural language deep parsing (provided by FRED [6]) whose output is a RDF/OWL representation of the entity definition; (iii) selection of candidate types (based on graph-pattern-based heuristics applied to FRED output); (iv) word-sense disambiguation of candidate types; and (v) type alignment to OntoWordNet [2], WordNet supersenses and to a subset of and DUL+DnS Ultralite. We refer to [3] for details about the design and the implementation of Tìpalo. In [3] we evaluated Tìpalo by extracting the types for a sample of 627 resources, while in this work we want to extract the ontology of Wikipedia by running Tìpalo on 3,769,926 DBpedia entities taken from the `dbpedia_long_abstracts_en` dataset of DBpedia, which include only entities having a Wikipedia abstract: this is a main constrain for applying our method.

4 The Natural Ontology of Wikipedia (ORA): results and discussion

The process described above has been run on a Mac Pro Quad Core Intel Xeon 2.8Ghz with 10Gb RAM and took 15 days (which can be easily reduced by parallelizing the activity on a cluster of machines with similar or more powerful

[7] http:schema.org

characteristics). The process resulted in 3,023,890 typed entities and associated taxonomies of types. Most of the missing results are due to the lack of matching Tìpalo heuristics, which means that by improving Tìpalo we will improve coverage (this is part of our current work). The resulting ontology includes 585,474 distinct classes organized in a taxonomy with 396,375 `rdfs:subClassOf` axioms; 25,480 if these classes are aligned through `owl:equivalentClass` axioms to 20,662 OntoWordNet synsets by means of a word-sense disambiguation process. The difference between the number of disambiguated classes (25,480) and the number of identified synsets (20,662) means that there are at least 4,818 synonym classes in the ontology. We expect the number of actual synonyms to be greater. Hence, we are planning to investigate some sense-similarity-based metric in order to reduce the number of distinct classes in the ontology by merging synonyms or at least providing explicit similarity relations with confidence scores between classes.

In order to prevent polysemy deriving from merging classes with same names but aligned to different synsets, it has been adopted a criterion of uniqueness for the generation of the URIs of these classes. For example, let us consider the entity `dbpedia:The_Marriage_of_Heaven_and_Hell`[8]. For this entity Tìpalo generates the following RDF:

```
dbpedia:The_Marriage_of_Heaven_and_Hell
    a  fred:Book .
fred:Book
    owl:equivalentClass  wn30-instance:synset-book-noun-2 .
```

Similarly, for the entity `dbpedia:Book_of_Revelation`[9] Tìpalo generates the following RDF:

```
dbpedia:Book_of_Revelation
    a  fred:CanonicalBook .
fred:CanonicalBook
    rdfs:subClassOf  fred:Book .
fred:Book
    owl:equivalentClass  wn30-instance:synset-book-noun-10 .
```

The two `fred:Book` classes refers to two distinct concepts. Hence, they cannot be merged during the generation of the ontology. We solve this by appending the ID of the closest synset in the taxonomy to the URI of the new generated classes: this approach guarantees to prevent polysemy and to identify synonymity at the same time. Finally, all the classes aligned to OntoWordNet have been also aligned to WordNet supersenses and a subset of DOLCE+DnS Ultra Lite classes by means of `rdfs:subClassOf` axioms. The following example shows a sample of the ontology which has been derived by typing the two entities used as examples previously:

[8] The definition of `dbpedia:The_Marriage_of_Heaven_and_Hell` is: *"The Marriage of Heaven and Hell is one of William Blake's books."*

[9] The definition of `dbpedia:Book_of_Revelation` is: textit"The Book of Revelation is the last canonical book of the New Testament in the Christian Bible."

```
dbpedia:The_Marriage_of_Heaven_and_Hell
    a  fred:Book_102870092 .
dbpedia:Book_of_Revelation
    a  fred:CanonicalBook_106394865 .
fred:CanonicalBook_106394865
    rdfs:subClassOf  fred:Book_106394865 ;
    rdfs:label  "Canonical Book"@en-US .
fred:Book_102870092
    owl:equivalentClass  wn30-instance:synset-book-noun-2 ;
    rdfs:label  "Book"@en-US .
fred:Book_106394865
    owl:equivalentClass  wn30-instance:synset-book-noun-10 ;
    rdfs:subClassOf  wn30-instance:supersense-noun_communication ,
                     d0:InformationEntity ;
    rdfs:label  "Book"@en-US .
```

Conclusion. The main result of this work is the Natural Ontology of Wikipedia (ORA): an ontology that reflects the richness of terms used and agreed by the crowds for defining entities in Wikipedia. All produced datasets are available for download[10]. We claim that this ontology provides an important resource that can be used as alternative or complement for YAGO and DBPO, and that it can enable more accurate usage of DBpedia in Semantic Web based applications such as: mash-up tools, recommendation systems, and exploratory search tools (see for example Aemoo [5]), etc. Currently, we are working at refining ORA and to align it to DBPO and YAGO.

References

1. P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* Springer, 2006.
2. A. Gangemi, R. Navigli, and P. Velardi. The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In *in WordNet, Meersman*, pages 3–7. Springer, 2003.
3. A. Gangemi, A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini. Automatic Typing of DBpedia Entities. In *International Semantic Web Conference (1)*, volume 7649 of *Lecture Notes in Computer Science*, pages 65–81. Springer, 2012.
4. J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, 2009.
5. A. G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, and P. Ciancarini. Aemoo: Exploring knowledge on the web. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 272–275. ACM, 2013.
6. V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *Knowledge Engineering and Knowledge Management (EKAW 2012)*, pages 114–129. Springer, 2012.
7. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, pages 697–706, New York, NY, USA, 2007. ACM Press.

[10] http://stlab.istc.cnr.it/stlab/ORA