WikiTaxonomy: A Large Scale Knowledge Resource

Simone Paolo Ponzetto[1] and **Michael Strube**[1]

Abstract. We present a taxonomy automatically generated from the system of categories in Wikipedia. Categories in the resource are identified as either *classes* or *instances* and included in a large subsumption, i.e. *isa*, hierarchy. The taxonomy is made available in RDFS format to the research community, e.g. for direct use within AI applications or to bootstrap the process of manual ontology creation.

1 INTRODUCTION

Advances in the development of knowledge intensive AI systems crucially depend on the availability of large coverage, machine readable knowledge sources. While tremendous progress in AI has been made in the last decades by investigating data-driven inference methods, we believe that further advancement ultimately depends also on the *free access to large repositories of structured knowledge* on which these inference techniques can be applied. In this article we approach the problem by using Wikipedia. We present methods for deriving a large coverage taxonomy of classes and instances from the network of categories in Wikipedia and present the RDF Schema we make freely available to the research community.

2 METHODS

We apply in sequence the methods described in Ponzetto & Strube [8] and Zirn et al. [13] in order to generate a semantic network from the system of categories in Wikipedia.

1. We label the relations between category pairs as *isa* and *notisa*. This way the category network, which *per-se* is merely a hierarchical thematic categorization of the topics of articles, is transformed into a subsumption hierarchy with a well-defined semantics.
2. We classify categories as either *classes* or *instances* in order to distinguish between *isa* subsumption and *instance-of* relations.

2.1 Deriving a taxonomy from Wikipedia

In [8] we presented a set of lightweight heuristics for distinguishing between *isa* and *notisa* links in the Wikipedia category network.

Syntax-based methods label category links based on string matching of syntactic components of the category labels. They use a full syntactic parse of the category labels to check whether category label pairs share the same lexical head[2] (*head matching*) or the head of a category label occurs as a modifier in another one (*modifier matching*).

Connectivity-based methods reason on the structure and connectivity of the categorization network. *Instance categorization* applies the method from [10] to identify instances from Wikipedia pages to those categories referring to the same entities as the pages. *Redundant categorization* labels category pairs as in an *isa* relation by looking for directly connected categories redundantly having a page in common.

Lexico-syntactic based methods use lexico-syntactic patterns applied to large text corpora (e.g. Wikipedia itself) to identify *isa* [4] and *part-of* relations [2], the latter providing evidence that the relation is not an *isa* relation. A majority voting scheme based on the number of hits for each set of patterns is used to decide whether the relation is *isa* or not.

Inference-based methods propagate the previously found relations based on the properties of multiple inheritance and transitivity of the *isa* relation.

These methods generate 105,418 *isa* links from a network of 127,325 categories and 267,707 links. We achieve a score of 87.9 balanced F-measure when evaluating the taxonomy against the subset of ResearchCyc [6] in which the categories can be mapped to.

2.2 Distinguishing between classes and instances

Zirn et al. [13] go one step forward from [8] and classify categories as *instances* or *classes*. This step yields a taxonomy with finer grained semantics, and it is necessary since the network contains many categories whose reference is an entity, e.g. the MICROSOFT category[3], rather than a property of a set of individuals, e.g. MULTINATIONAL COMPANIES. Similarly to [8], they devise a set of heuristics on which to decide the reference type of a category label and combine the best performing methods for each class into a voting scheme. Given a category c with label l, c is classified as either an *instance* or a *class* by the first satisfied criterion.

1. **Page & Plural**: if no page titled l exists and the lexical head of l is plural, then c is a *class*.
2. **Capitalization & NER**: else if l is capitalized and has been recognized by a Named Entity Recognizer as a named entity, then c is an *instance*.
3. **Page**: else if no page titled l exists, then c is a *class*.
4. **Plural**: else if the head of l is plural, then c is a *class*.
5. **Structure**: else if c has no sub-category, then it is a *class*.
6. **Capitalization**: else if l is capitalized, then c is an *instance*.
7. **Default**: else c is a *class*.

Using the same category network from [8] this pipeline of heuristics is shown to classify 111,652 *class* and 15,472 *instance* categories with an accuracy of 84.5% when evaluated against ResearchCyc.

[1] EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg, Germany. Website: http://www.eml-research.de/nlp

[2] The head of a phrase is the word that determines the syntactic type of the overall phrase of which it is a member. In the case of category labels, it is the main noun of the label, e.g. the noun Scientists for the category label SCIENTISTS WHO COMMITTED SUICIDE.

[3] We use Sans Serif for words and queries, CAPITALS for Wikipedia pages and SMALL CAPS for Wikipedia categories.

```
<rdf:Description rdf:about="http://www.eml-research.de/WikipediaOntology/Class#_1268">
  <rdfs:subClassOf rdf:resource="http://www.eml-research.de/WikipediaOntology/Class#_2419"/>
  <rdfs:comment>http://en.wikipedia.org/Wiki/Category:Multinational_companies</rdfs:comment>
  <rdfs:label>Multinational_companies</rdfs:label>
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:Description>

<rdf:Description rdf:about="http://www.eml-research.de/WikipediaOntology/Individual#:_36">
  <rdfs:comment>http://en.wikipedia.org/Wiki/Category:Microsoft</rdfs:comment>
  <rdfs:label>Microsoft</rdfs:label>
  <rdf:type rdf:resource="http://www.eml-research.de/WikipediaOntology/Class#_1268"/>
</rdf:Description>
```

Figure 1. Fragment of WikiTaxonomy in RDFS format. Individuals are linked to the class they are instances of using the `rdf:type` predicate.

3 WIKITAXONOMY

We applied the methods from [8] and [13] using the English Wikipedia database dump from 25 September 2006. The extracted taxonomy was converted into RDF Schema [3, RDFS] using the Jena Semantic Web Framework[4]. RDFS has a very limited semantics and serves mostly as foundation for other Semantic Web languages. Nevertheless it suffices in the present scenario of data exchange where we have only a set of classes in a hierarchical relation. RDFS in addition provides compatibility with free ontology editors such as Protégé [5] for visualization, additional manual editing or conversion to richer knowledge representation languages such as OWL [7]. Figure 1 shows a sample fragment of the WikiTaxonomy in RDFS format. In the RDFS data model Wikipedia categories are represented as *resources* (i.e. a list of `rdf:Description` elements) and the subsumption relation is modeled straightforwardly using the `rdfs:subClassOf` property. A human readable version of the name of the category is given via the `rdfs:label` property and a link to the on-line version of the corresponding page is provided using the `rdfs:comment` property. In order to distinguish between categories which are instances or classes we use the `rdf:type` predicate to state whether a resource is a class or an individual of a class. In addition, the distinction is also given in the resource identifier, i.e. the URI-reference.

4 RELATED WORK

Researchers working in information extraction have recently begun to use Wikipedia as a resource for automatically deriving structured semantic content. Suchanek et al. build the YAGO system [10] by merging WordNet and Wikipedia: the *isa* hierarchy of WordNet is populated with instances taken from Wikipedia pages. Auer et al. present the DBpedia system [1] which generates RDF statements by extracting the attribute-value pairs contained in the *infoboxes* of the Wikipedia pages (i.e. the tables summarizing the most important attributes of the entity referred by the page), e.g. the pair `capital=[[Berlin]]` from the GERMANY page. Wu & Weld show in [11] how to augment Wikipedia with automatically extracted information. They propose to 'autonomously semantify' Wikipedia by (1) extracting new facts from its text via a cascade of Conditional Random Field models; (2) adding new hyperlinks to the articles' text by finding the target articles nouns refer to. Wu & Weld's Kylin Ontology Generator (KOG) [12] is the work closer to ours. Their system builds a subsumption hierarchy of classes by combining Wikipedia infoboxes with WordNet using statistical-relational learning. Each infobox template, e.g. `Infobox Country` for countries, represents a class and the slots of the template are considered as the attributes of the class. KOG uses Markov Logic Networks [9] in order to jointly predict both the subsumption relation between classes and their mapping to WordNet. While KOG represents a theoretically sounder methodology than [8] and [13], the lightweight heuristics from the latters are straightforward to implement and show that, when given high quality semi-structured input as in the case of Wikipedia, large coverage semantic networks can be generated by using simple heuristics which capture the conventions governing its public editorial base.

[4] `http://jena.sourceforge.net`

ACKNOWLEDGEMENTS

This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant (09.003.2004).

REFERENCES

[1] Sören Auer, Christian Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives, 'DBpedia: A nucleus for a Web of open data', in *Proc. of ISWC 2007 + ASWC 2007*, pp. 722–735, (2007).

[2] Matthew Berland and Eugene Charniak, 'Finding parts in very large corpora', in *Proc. of ACL-99*, pp. 57–64, (1999).

[3] Dan Brickley and Ramanathan V. Guha, 'RDF vocabulary description language 1.0: RDF schema', Technical report, W3C, (2004). `http://www.w3.org/TR/rdf-schema`.

[4] Marti A. Hearst, 'Automatic acquisition of hyponyms from large text corpora', in *Proc. of COLING-92*, pp. 539–545, (1992).

[5] Holger Knublauch, Ray W. Fergerson, Natalya Fridman Noy, and Mark A. Musen, 'The Protégé OWL plugin: an open development environment for semantic web applications', in *Proc. of ISWC 2004*, pp. 229–243, (2004).

[6] Douglas B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*, Addison-Wesley, Reading, Mass., 1990.

[7] Peter F. Patel-Schneider, Patrick Hayes, and Ian Horrocks, 'OWL Web Ontology Language semantics and abstract syntax', Technical report, W3C, (2004). `http://www.w3.org/TR/owl-semantics`.

[8] Simone Paolo Ponzetto and Michael Strube, 'Deriving a large scale taxonomy from Wikipedia', in *Proc. of AAAI-07*, pp. 1440–1445, (2007).

[9] Matthew Richardson and Pedro Domingos, 'Markov logic networks', *Machine Learning*, **62**, 107–136, (2006).

[10] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum, 'YAGO: A core of semantic knowledge', in *Proc. of WWW-07*, pp. 697–706, (2007).

[11] Fei Wu and Daniel Weld, 'Automatically semantifying Wikipedia', in *Proc. of CIKM-07*, pp. 41–50, (2007).

[12] Fei Wu and Daniel Weld, 'Automatically refining the Wikipedia infobox ontology', in *Proc. of WWW-08*, (2008).

[13] Cäcilia Zirn, Vivi Nastase, and Michael Strube, 'Distinguishing between instances and classes in the Wikipedia taxonomy', in *Proc. of ESWC-08*, pp. 376–387, (2008).