Ontological Quality Control in Large-scale, Applied Ontology Matching

Catherine Legg, Samuel Sarjant
The University of Waikato, New Zealand

Email: clegg@waikato.ac.nz, sarjant@waikato.ac.nz

Abstract. To date, large-scale applied ontology mapping has relied greatly on label matching and other relatively simple syntactic features. In search of more holistic and accurate alignment, we offer a suite of partially overlapping ontology mapping heuristics which allows us to *hypothesise* matches and test them against the knowledge in our source ontology (OpenCyc). We thereby automatically align our source ontology with 55K concepts from Wikipedia with 93% accuracy.

1. Introduction

We have developed a method of specifically *ontological* quality control in ontology mapping which combines a suite of partially overlapping mapping heuristics with common-sense knowledge in OpenCyc. Our approach differs from previous largely label-matching approaches (Suchanek et al, 2008, Ponzetto and Navigli, 2009) in its use of knowledge, and also from previous knowledge-based approaches (Shvaiko and Euzenat, 2005, Sabou et al, 2006), in treating potential matches as *hypotheses*, and testing them more iteratively and open-endedly than previously accomplished.

2. Iterative Mapping Process

Concept to Wikipedia article mapping is governed by a *priority queue* which iteratively evaluates potential mappings ordered via continuously updated weightings. The process begins with concept-to-article mappings (**Table 1**), then verifies these using article-to-concept heuristics. The weight of each potential mapping is equal to the product of weights produced by the two sets of heuristics.

Table 1. Heuristics that map between source ontology concepts and Wikipedia articles.

Concept → Article	**Example**
TITLE MATCHING	`Batman-TheComicStrip` → {*Batman (comic strip)*:1.0}
SYNONYM MATCHING	`ComputerWorm` → {*Worm*:1.0, *Computer worm*:0.39, ... (+5 more)}
CONTEXT-RELATED SYNONYM MATCHING	`ComputerWorm` → {*Computer worm*:1.0, *Worm*:0.59,... (+4 more)}
Article → Concept	**Example**
TITLE MATCHING	*Dog* → {`Dog`:1.0, `HotDog`:1.0}
LABEL MATCHING	*Dog* → {`Dog`:1.0, `HotDog`:0.995, `CanineAnimal`:0.03, `CanineTooth`:0.03}

A final quality control measure is the 'consistency check' between information on concept and the mapped article. Most Wikipedia first sentences are conventionally structured as: '*X* is/was/are/were a/an/the *Y*', where *Y* is links to articles typically

representing appropriate classes. The mapping weight is multiplied by the proportion of assertions not rejected using OpenCyc's disjointness knowledge.

Example 1: *"**Bill Laswell** is an American [[bassist]], [[record producer|producer]] and [[record label]] owner."* Only three of the four assertions in this sentence are kept: `BillLaswell` is a `UnitedStatesPerson`, `BassGuitarist`, and `Producer`. `BillLaswell` cannot be a `RecordCompany` because OpenCyc knows a person cannot be a company.

Example 2: The concept `Basketball-Ball` initially maps as follows (*Basketball:*1.0, *Basketball (ball)*:0.95, *College basketball:*0.02). The second candidate is the correct one, as the first refers to the team sport. The algorithm attempts to map its first choice *Basketball* back to `Basketball-Ball`, which succeeds but also creates a new potential reverse mapping *Basketball* → `Basketball`. Consistency checking now tests "`Basketball-Ball` is a `TeamSport`", which fails, removing this potential mapping. The next highest reverse-mapping is *Basketball* → `Basketball`, which is found to be consistent, so a mapping is recorded for that. The process now backtracks to hypothesising the second-best option from the original list: *Basketball (ball)*:0.95, which also successfully reverse-maps and is consistent, creating a new (correct) mapping. It is worth emphasising how similar the two 'basketball concepts' are by standard semantic relatedness measures, and thus the subtlety our methods are capable of.

3. Results and Conclusions

The algorithm identified 54,987 mappings of OpenCyc concepts to Wikipedia articles. Applying manual analysis to a random 300 mappings, 266 were judged 'True' (88.5%), 21 'False' (7%) and 13 (4.3%) were assigned 'B' for 'Broader term' (the mapping was largely correct but one side generalised the other). Thus 93% of our mappings were either 'True' or highly related. Although YAGO reports 95% accuracy, what is being rated is not mapping joins between Wordnet and Wikipedia, but the truth of assertions in infoboxes. Although our efforts so far lack the scale of projects such as YAGO, we suggest they have a role to play in long-term development towards maximum accuracy in this field. We offer our results at: `http://bit.ly/10MlLjl`.

References

Euzenat, J. and Shvaiko, P. (2007). *Ontology Matching*. Springer-Verlag.

Ponzetto, S.P., and Navigli, R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia, *IJCAI 2009*, Pasadena, California, pp. 2083-2088.

Sabou, M., D'Aquin, M., Motta, E. (2006). Using the Semantic Web as Background Knowledge for Ontology Mapping, *OM-2006*, Athens, GA, USA.

Shvaiko, P. and Euzenat, J. (2005). A Survey of Schema-based Matching Approaches. *Journal on Data Semantics* **4**.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A Large Ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics* **6**(3), 203-217.