Flexible bootstrapping-based ontology alignment

Prateek Jain, Pascal Hitzler and Amit P. Sheth,

Kno.e.sis Center, Dept. Of Computer Science & Engineering, Wright State University

http://knoesis.wright.edu

Problem Analysis

Ontology matching systems traditionally rely on external sources for performing a match between various elements such as classes, properties and instances. These external resources can, for example, be any of the following.

Linguistic Resources: Thesauri such as WordNet or lexicons are utilized to match the words and their linguistic relationships to other terms.

Upper Level Ontologies: Upper level ontologies like SUMO or DOLCE have been utilized in the past for matching ontological concepts.

Reusing Existing Alignments: By utilizing the existing alignments computed previously, new mappings between ontologies can be computed by utilizing a reasoner or any other tool which can consume these mappings.

ISSUES WITH THESE APPROACHES

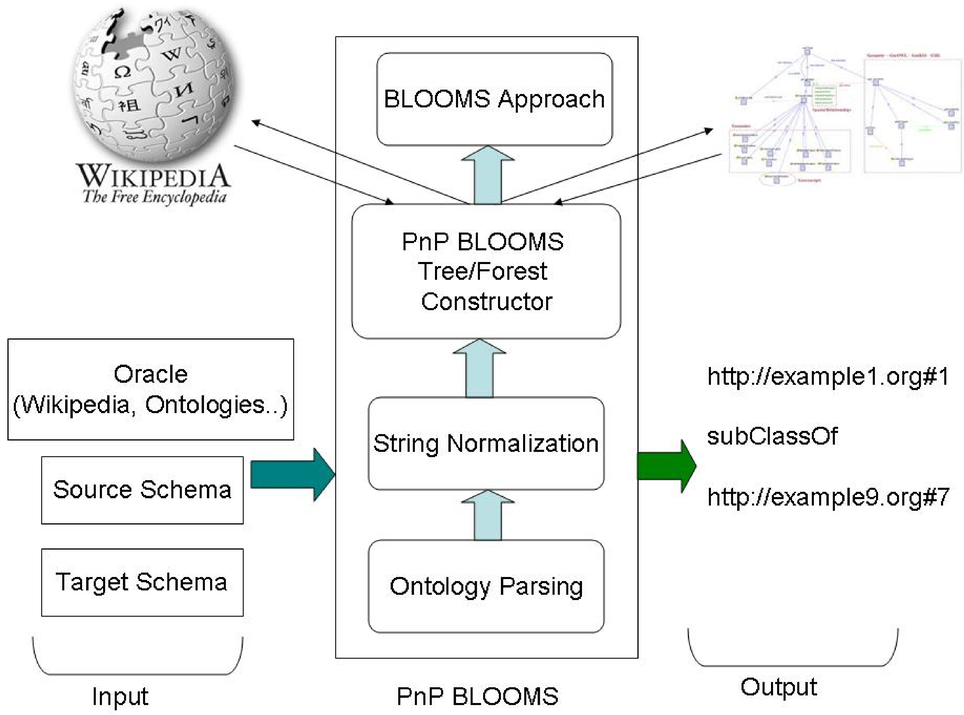
Ontologies are in Other Languages: Tools relying on lexicons and other language dependent tools do not work well on ontologies consisting of terms in other languages. We witnessed this problem first hand while evaluating BLOOMS on the oriented matching track of the OAEI 2009 initiative. While BLOOMS had an f-measure of 1 and 0.91 for the 1XX and 3XX evaluations, it dipped to 0.53 for the 2XX track

Ontologies belong to narrow domain: In the Geonames ontology, the concept 'Code' denotes codes representing geographical features such as county. However, in case of a biomedical ontology, the same phrase may represent genetic code. Thus, for tools which are developed or tuned for certain domains, their usage outside of their domain can result in unsatisfactory performance.

Large-scale nature of the ontologies: The Linked Open Data (LOD) cloud has allowed for the manifestation of large domain specific ontologies. Many of the real world ontologies used in LOD have over 100 concepts. Thus, ontology alignment tools need to be able to scale to the levels in order to be of any practical use.

Requirement for Domain Specific Alignment: While it is expected to assert equivalence between concept 'Code' appearing in bio-medical ontologies, while equivalence between 'Code' in bio-medical ontology with 'Code' in a commercial billing ontology is irrelevant and wrong.

Some of these issues have been concisely summarized in [2].



Evaluation

Comparison of BLOOMS and PnP BLOOMS on French ontologies of the Benchmark track of OAEI 2009.

Ontology	f-measure BLOOMS	f-measure PnP BLOOMS
206	0.53	0.76
207	0.58	0.77
210	0.50	0.72
Average	0.54	0.75

Our Solution

In a nutshell, BLOOMS constructs a forest (i.e., a set of trees) T_C (which we call the BLOOMS forest for C) for each matching candidate class name C – in the original approach of BLOOMS [1], this roughly corresponds to a selection of supercategories of the class name in the sense of the Wikipedia class hierarchy. Comparison of the forests T_C and T_B for matching candidate classes C and B then yields a decision whether or not (and with which of the candidate relations) C and B should be aligned.

1.Pre-processing of the input ontologies in order to(i) remove property restrictions, individuals, and properties, and to (ii) tokenize compound class names to obtain a list of all simple words contained within them, with stop words removed.

2.Parsing of the Auxiliary Datasources: PnP BLOOMS supports data sources namely Wikipedia's in other languages and RDFS/OWL ontologies. The parsers of BLOOMS can be utilized as is for utilizing other language Wikipedia, modifications has been made for ontologies like SUMO, Cyc due to use of named relationships.

3.Construction of the PnP BLOOMS forest TC for each class name C, using in-formation from the auxiliary data source selected by the user.

4.Comparison of constructed PnP BLOOMS forests, which yields decisions which class names are to be aligned.

5. Post-processing of the results with the help of the Alignment API and a reasoner.

Acknowledgements

The work is funded primarily by NSF Award:IIS-0842129, titled "III-SGER: Spatio-Temporal-Thematic Queries of Semantic Web Data: a Study of Expressivity and Efficiency". Pascal Hitzler acknowledges support by the Wright State University Research Council.

References

1.Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma, Peter Z. Yeh: Ontology Alignment for Linked Open Data. Proceedings of the 9th International Semantic Web Conference 2010, Shanghai, China, November 7th-11th, 2010. Pages 402-417 (To appear).

2.Pavel Shvaiko and Jerome Euzenat: Ten challenges for ontology matching. On the Move to Meaningful Internet Systems: OTM 2008 (2008) 1164–1182.