CompositeMatch: Detecting N-Ary Matches in Ontology Alignment*

Kelly Moran[1], Kajal Claypool[1], and Benjamin Hescott[2]

[1] MIT Lincoln Laboratory
{kmoran,claypool}@ll.mit.edu
[2] Tufts University
hescott@cs.tufts.edu

Abstract. The field of ontology alignment still contains numerous unresolved problems, one of which is the accurate identification of composite matches. In this work, we present a context-sensitive ontology alignment algorithm, CompositeMatch, that identifies these matches, along with the typical one-to-one matches, by looking more broadly at the information that a concept's relationships confer. We show that our algorithm can identify composite matches with greater confidence than current tools.

1 Introduction

Numerous ontology alignment algorithms have been developed during recent years to handle the growing challenge of aligning ontologies. While numerous advances have been made, most ontology alignment algorithms are still somewhat inadequate at identifying complex relationship-based matches, a la composite matches. A composite match is defined as a match between a concept c_0 in ontology O and multiple concepts $c_0'\ldots c_n'$ in ontology O'; the reverse; or alternatively between multiple concepts in O and multiple concepts in O'. For example, if an ontology O contains the concept *Name* and an ontology O' contains concepts *FirstName* and *LastName*, then *Name* matches neither *FirstName* nor *LastName* but the composite of the two. In this work, we present *CompositeMatch*, an algorithm that combines linguistic and contextual approaches for semi-automatically discovering both one-to-one and composite matches.

2 The CompositeMatch Algorithm

CompositeMatch is a three-pass algorithm that operates on two input ontologies and produces an alignment file containing matches between them. The **first phase** performs a linguistic match between the ontologies' concepts. The linguistic match assigns a normalized similarity score between 0 and 1.0 to each pair.

In the **second phase**, the most uncertain pairs– collectively referred to as the *grey zone*– are judged on contextual criteria to determine whether they should be accepted as viable matches. The grey zone consists of all conflicting matches– matches that contain the same concept, rendering unclear which match is the true match for the concept– plus matches with a similarity score between the upper and lower thresholds set prior to execution. The second phase defines two rules that serve as a filtering process to increase the scores of contextually similar matches.

The **third and final phase** is a post-processing phase that scours the matches for possible composite matches, looking again at contextual criteria for any indicative information, before outputting the final set of matches to the user.
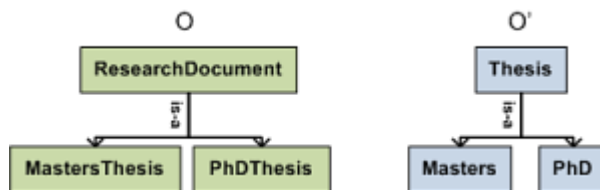

Fig. 1. An m:n composite match

Example 1. Consider the case shown in Figure 1. The first phase finds some similarity between the concept *MastersThesis* in ontology O and concepts *Masters* and *Thesis* in ontology O', as well as *PhDThesis* in O and *PhD* and *Thesis* in O'. Phase 2 compares the parents of these pairs but does not find sufficient contextual similarity to augment the strengths of any of them. In Phase 3, the algorithm finds that the conflicting matches identified earlier result in composite concepts: the concepts *MastersThesis* and *PhDThesis* from O form a composite concept that matches the composite concept between *Thesis*, *Masters*, and *PhD* in O'.

3   Preliminary Results

We evaluated the performance of CompositeMatch on two tests, the first being the benchmark test from the Ontology Alignment Evaluation Initiative (OAEI) 2008. Because the OAEI benchmark does not account for composite matches, we created a second test: a set of six ontologies, each a modified version of the OAEI benchmark base case, into which composite matches were injected.

We compared the results of CompositeMatch on the OAEI 2008 benchmark to those of a high-performing OAEI 2008 entrant, RiMOM. The mean performance of CompositeMatch on the tests is a precision of .926 and a recall of .557. RiMOM has an overall precision of .939 and a recall of .802. We also considered the subset of all tests not including random or foreign language labels as we believe this subset is a better indicator of how CompositeMatch performs on its intended data sets. CompositeMatch achieves a significantly higher precision and recall on this subset of .996 and .896 respectively, and RiMOM increases slightly to .965 and .967.

While further evaluation is needed, our preliminary results indicate that CompositeMatch correctly identifies each composite match, achieving both a precision and a recall of 1.0, while RiMOM identifies none.