Using Semantic Similarity in Ontology Alignment

Valerie Cross and Xueheng Hu

Computer Science and Software Engineering Department,
Miami University, Oxford, OH 45056
crossv@muohio.edu

Abstract. Many approaches to measure the similarity between concepts that exist in two different ontologies are used in the matchers of ontology alignment systems. These matchers belong to various categories depending on the context of the similarity measurement, such as lexical, structural, or extensional matchers. Although OA systems have used various forms of similarity measures along with some background knowledge sources, not many have incorporated the use of semantic similarity measures. This paper first reviews the use of semantic similarity in current OA systems, presents a unique application of such measures to assess the semantic alignment quality (SAQ) of OA systems and reports on the results of a study done using SAQ measures on the OAEI 2010 results from the anatomy track

Keywords: Semantic similarity, ontological similarity, ontology alignment, information content, semantic alignment quality.

1 Introduction

Ontology alignment (OA) research has typically concentrated on finding equivalence relationships between different concepts in different ontologies. The result of the OA process is typically a set of mappings between concepts from two different ontologies with a confidence value in [0, 1] for each mapping. OA techniques vary greatly depending on the features used to determine the mapping, i.e., the schema, its instances, etc. and the background knowledge sources used such as vocabularies or other ontologies, already existing alignments, free text, etc. Another term semantic matching has been used to describe the process when not only equivalence relations but also generalization and specialization relations are determined [1].

Early OA work focused on using string edit distances between the concept labels and the overall structure of the ontologies. Even in the same domain a wide variance in the terminology and the structuring of the concepts may still exist. Much research has worked on handling these wide variations in ontologies. GLUE [2] was one of the first to combine several different learners (similar to what is currently called a matcher) to establish mappings. A learner using instance information and one using a concept's complete list of ancestor concepts from the ontology root to the concept itself were combined to determine concept similarity between the two ontologies.

The basis for many matchers in OA systems can be found in [3] where Tversky's parameterized ratio model of similarity [4] is used with various features of concepts. Many of these similarity measures have been adapted for use in matchers in various categories depending on the context of the similarity measurement, such as lexical, structural, or extensional matchers [5]. General ontologies such as WordNet [6] have been used to find synonyms for differing concept string labels. The OLA system [7] calculates lexical similarity between two concepts by looking up their names in WordNet to find the synonyms for each concept. It does a string-based match between the pairs of synonyms and an aggregation on the resulting string similarities. RiMoM [8][9] incorporates the UMLS Metathesaurus [10] to align biomedical domain ontologies and general background knowledge sources such as Wiki to align common knowledge ontologies. More recent OA systems incorporate background knowledge sources to improve the OA process. AgreementMaker [11][12][13] extends its string-based matchers by integrating lexicons. The WordNet Lexicon is built to incorporate the synonym and definition annotations found in the ontologies themselves and then augments these with any non-duplicated synonyms and definitions existing in WordNet that correspond to those in the ontologies being aligned. The string-based matchers then work not only on the specific concept labels but also on the corresponding synonyms in the WordNet Lexicon. ASMOV [14] optionally permits a thesaurus to be used, either the UMLS Metathesaurus or WordNet, to calculate the lexical similarities between each pair of concepts, properties and individuals.

Although various forms of similarity measures are used in OA systems, only a few have incorporated semantic similarity in the OA process. This paper examines the use of semantic similarity for the evaluation of a mapping set produced by an OA system. Traditional OA evaluation strategies generally depend on a reference alignment considered to be a correct and complete set of mappings between the two ontologies and determined by a domain expert. Given a reference alignment, the quality of an OA system is evaluated with the three standard criteria: precision, recall, and f-measure. This evaluation approach has two obvious disadvantages. First, the reliability of the evaluation is directly determined by the quality of the reference alignment. For example, the reference alignment may only capture limited information of the related domain and be incomplete so OA system mappings might be correct but not found in the reference alignment. Second, in many practical cases, a reference alignment may not be available or requires too much effort to create.

This research proposes using semantic similarity measures for OA evaluation purposes, that is, a semantic alignment quality (SAQ) measure for use in addition to or in place of the standard three measures when a reference alignment is not available. The SAQ measure assesses the quality of a pair of mappings by comparing the semantic similarity between two concepts in the source ontology with the semantic similarity between the two target concepts they are mapped to. This process is performed on all pairs of mappings in the OA result to determine an overall SAQ.

First Section 2 reviews semantic similarity measures and provides examples of their use with background knowledge in current OA systems. Section 3 describes the SAQ measure. Section 4 presents the details and analysis of the experiments conducted using a wide variety of semantic similarity measures within the SAQ measure on the OAEI 2010 anatomy track ontologies. Section 5 summarizes the research and outlines plans for future research.

2 Semantic Similarity in OA

In ontology research, semantic similarity measurement is typically used to assess the similarity between concepts within an ontology. Cross-ontological similarity measures [15], i.e., ones that measure the similarity between concepts in different ontologies based on establishing association links between the concepts have been proposed. Another approach develops semantic similarity measures between concepts based on the description logic definition of the concepts. These approaches vary depending on what sets the similarity is measured such as instance sets [16], characteristic sets [17], or model sets [18]. Future research should investigate the usefulness of the cross-ontological and DL based semantic similarity measures in OA evaluations. The focus here, however, is semantic similarity measured within one ontology and using the subsumption relationship. Such semantic similarity measures are currently being used in OA systems with background knowledge sources. These semantic similarity measures were first divided into two main categories: path or distance-based and information content based. Later, set-based semantic similarity measures followed Tversky's parameterized ratio model of similarity [4]. A brief overview of these three categories, example measures, and references to some OA systems using such measures is provided [19].

The *path-based similarity measures* or edge-counting similarity measures rely on the distance between two concepts. This distance is a count of the number of edges on the path or a count of the number of nodes in the path linking the two concepts. Some approaches assign different weights to edges or use different conversions and normalizations of Rada's distance metric [20] into a similarity measure. For example, Leacock and Chodorow [21] converted Rada's distance metric into a path-based semantic similarity as follows:

$$sim_{LC} = -\log(\min_p[len(p(c1,c2))]/2D)$$

where D is the depth of the ontology that contains c1 and c2. It basically normalizes Rada's distance measure *len(p(c1,c2))* using D and converts it to similarity by using the negative logarithm. An early OA system iMapper [22] uses a simple path based semantic distance between two terms *x* and *y* found in WordNet. If they belong to the same WordNet synset, then the path distance is 1. Otherwise, the path length is determined by first finding the paths from each sense of *x* to each sense of *y*, counting the number of nodes in each path between the two senses, and using the minimum count of nodes for the semantic distance. Note that path length is determined by the number of nodes rather than number of edges in the path.

The Wu and Palmer measure [23] calculates similarity using the distance from the root to the common subsumer of c1 and c2. The formula is:

$$sim_{WP}(c1, c2) = 2\times \frac{len(root,c3)}{len(c1,c3) + len(c2,c3) + 2\times len(root,c3)}$$

where c3 is the common subsumer of c1 and c2. In the case that c1 and c2 have multiple common subsumers, c3 is typically assumed to be the lowest, i.e., the one with the greatest distance from the root. For this research, c3 is selected as the one

that minimizes the path distance between c1 and c2 since in a well-designed ontology, this c3 should also be the lowest one. OLA [7] uses a measure similar to the Wu-Palmer measure with the WordNet ontology. ASMOV [14] use the Wu-Palmer semantic similarity on the XML data type hierarchy for properties, when the ranges of two data type properties are being compared.

Information content (IC) based measures use a measure of how specific a concept is in a given ontology. The more specific a concept is the higher its IC. The more general a concept is the lower its IC. Originally, IC uses an external resource such as an associated corpus [24]. The corpus-based IC measure for concept *c* is given as

$$IC_{corpus}(c) = -\log p(c)$$

where the value p(c) is the probability of the concept determined using the frequency count of the concept, i.e. the number of occurrences within the corpus of all words representing the concept and includes the total frequencies of all its children concepts.

The ontology-based IC [25] uses the ontology structure itself [25] and is defined as

$$IC_{ont}(c) = \log \frac{(num\_desc(c)+1)}{max_{ont}} / \log \frac{1}{max_{ont}} = 1 - \frac{\log(num\_desc(c)+1)}{\log(max_{ont})}$$

where num_desc(c) is the number of descendants for concept *c* and max_{ont} is the maximum number of concepts in the ontology. This IC measure is normalized such that the information content values are in [0...1]. IC_{ont} has maximum value 1 for the leaf concepts and decreases until the value is 0 for the root concept of the ontology.

The first IC based ontological similarity measure was proposed by Resnik [24] as

$$sim_{RES}(c1,c2) = max_{S(c1,c2)} [IC_{corpus}(c)]$$

where S(*c1*,*c2*) is the set of concepts that subsume both *c1* and *c2*.

Lin [26] defined a measure that uses not only the shared information between the two concepts but also the separate information content of the two concepts:

$$sim_{Lin}(c1,c2) = \frac{2 \times IC(c3)}{IC(c1)+IC(c2)}$$

where *c3* is the subsuming concept with the most information content. ASMOV [14] uses the Lin measure to assess the semantic similarity between two labels in a thesaurus which is either WordNet or UMLS. UFOme [27] uses the Lin measure in its WordNet matcher to determine the semantic similarity between synsets found in WordNet when the concepts being mapped do not share the same synset in WordNet.

Jiang and Conrath [25] define another distance measure integrating path and information content based measures. The distance is based on totaling up their separate IC and subtracting out twice the IC of their most informative subsumer.

$$dist_{JC}(c1, c2) = IC(c1) + IC(c2) - 2 \times IC(c3)$$

so that the remaining IC indicates the distance between them. If no IC is left, i.e., 0, the two concepts are the same. This distance measure can be converted to similarity. Several approaches have been proposed. In [25], the following formula is used

$$sim_{JC}(c1, c2) = 1 - (IC(c1) + IC(c2) - 2 \times IC(c3)) \times 0.5.$$

Set-based semantic similarity measures use Tversky's parameterized ratio model [4]:

$$S_{Tverksy}(X, Y) = \frac{f(X \cap Y)}{f(X \cap Y) + \alpha f(X - Y) + \beta f(Y - X)}$$

where f is an evaluation measure on sets. The α and β permit variations on the similarity measure. Here f is defined as fuzzy set cardinality. Itbparallels set cardinality. The only difference is an element's degree of membership in the fuzzy set is added in instead of simply a 1 for the element. A concept's IC value is used as its membership degree. Fuzzy set cardinality of a set of concepts is the sum of each concept's IC in the set. A wide variety of fuzzy set similarity measures are based on the Tversky model [29]. One can view a concept in an ontology as an object with a set of features or a related set. If $\alpha = \beta = 1$, S becomes the fuzzy set Jaccard index:

$$S_{Jaccard}(c1, c2) = \frac{\sum_{c \in (relatedSet(c1) \cap relatedSet(c2))} IC(c)}{\sum_{c \in (relatedSet(c1) \cup relatedSet(c2))} IC(c)}$$

If $\alpha = \beta = 0.5$, S becomes the Dice coefficient. If $\alpha = 1$, $\beta = 0$, S becomes the degree of inclusion of the related set for concept c1 within the related set for concept c2.

Many different sets can be a related set of a concept c. In this research the upset which is the ancestor set of c in addition to the concept c itself, the downset which is the descendant set of c in addition to the concept c itself, and the hourglass which is the union of the upset and downset of a concept [30] are used. Other feature sets for a concept are entirely possible such as the neighborhood set of a concept where neighbors can be based on other relationship types besides the is-a and a parameter may be used to determine how wide the neighborhood is from the concept [3].

3 Semantic Alignment Quality

The SAQ measure determines how well each pair of mappings, (s_i, t_i) and (s_j, t_j) maintains the same semantic similarity between the corresponding concepts in each ontology. A good pair of mappings should result in $|sim(s_i,s_j) - sim(t_i,t_j)|$ being close to 0. In [30] a similar approach is taken based on ordered concept lattice theory and proposes two new distance measures. The upper cardinality distance d_u and lower cardinality d_l between concepts a and b are defined as

$$d_u(a, b) = |upset(a)| + |upset(b)| - 2*\max_{c\text{-join}}[\,|upset(c)|]$$
$$d_l(a, b) = |downset(a)| + |downset(b)| - 2*\max_{c\text{-meet}}[|downset(c)|]$$

where c-join is the join concept, an ancestor concept shared between a and b in the lattice with no other concept less than it in the concept lattice. The lower cardinality distance between concepts a and b is defined similarly to upper except the upset is

replaced by downset and the join is replaced by the meet, i.e., c-meet is a meet concept, a descendent concept shared between a and b in the lattice with no other concept greater than it in the concept. In [30], only results with d_l are reported using the OAEI 2009 anatomy track. The experiments reported here are performed with a wide variety of semantic similarity measures more familiar to the ontology research community than d_l. The d_l measure, however, is also used within SAQ for comparison purposes. The experimental results also show some considerations on using semantic similarity measures to evaluate OA results not examined in [30].

To more clearly explain the approach, assume the set of mappings M = {(s_i, t_i) | $s_i \in O_s$, $t_i \in O_t$, and s_i maps to t_i in the OA result set}. To measure the similarity difference for two mappings m_i and m_j, the following formula is used:

$$simDiff(m_i, m_j) = | sim(s_i, s_j) - sim(t_i, t_j) |$$

with s_i and s_j being source anchors and t_i and t_j being target anchors such that $m_i = (s_i, t_i)$ and $m_j = (s_j, t_j)$ in M. The overall difference of semantic similarity for source anchor pairs and target anchor pairs is calculated as

$$simDiff_{overall}(M) = \sum_{m_i, m_j \in M} simDiff(m_i, m_j)$$

and the average difference is calculated over all (m_i, m_j) pairs in M where $i \neq j$ as

$$simDiff_{average}(M) = \frac{simDiff_{overall}(M)}{C_2^N} \quad \text{where N = |M|.}$$

The denominator is the number of combinations of N mappings taken two at a time.

The SAQ measure is 1- $simDiff_{average}$. The closer SAQ is to 1, then the smaller the semantic similarity difference is over all the pairs of mappings in the alignment. More specifically, the alignment results of high quality are expected to produce small values for the $simDiff_{average}$. Here 'small values' means being close to zero or no greater than a predefined threshold. This threshold can be derived through experimentation with existing reference alignments that are believed to have high quality.

Notice that the SAQ can have any semantic similarity measures substituted for *sim*. The experiments reported in the next section used the lower cardinality distance, the two path based measures, the three IC based measures and 9 variations of the set-based similarity measures resulting from the three standard Tverskey set-based similarity measures paired with the three different related sets, the downset, the upset and the hourglass for a concept. The experiments investigate the performance differences of these semantic similarity measures in SAQ and if the notion of SAQ corresponds with the standard performance measures used to evaluate OA results.

4 Experimenting with SAQ and the OAEI 2010 Results

The ontology alignment evaluation initiative (OAEI) [31] conducts yearly competitions that include the most up-to-date OA systems. These systems and their algorithms are evaluated using the same set of test cases so that performance comparisons can be made by those interested in using them. The OAEI 2010 campaign supported four tracks: anatomy, benchmark, conference, and directory.

Each track is specialized for different purposes. The experiments reported in this section focus on the anatomy track since the anatomy track uses two real-world ontologies from the biomedical domain, the NCIT human anatomy (HA) ontology and the mouse anatomy ontology (MA) which are considerably larger and also produce many more mappings than those of the other tracks. The reference alignment between the two ontologies is readily available and consists of 1520 mappings. The precision, recall and f-measure of each OA system that participated in the anatomy track are also available. Finally, the concept lattice lower distance measure research in [30] also used the anatomy track of the 2009 OAEI.

Table 1 lists the OA systems alphabetically along with the number of mappings produced and their performance measures on the anatomy track's first subtask which is to produce the best mappings possible emphasizing the f-measure.

Table 1. OA Systems OAEI 2010 Anatomy Track Precision (P), Recall (R), F-Measure (F)

OA Systems	# of mappings	P	R	F
AgrMaker	1436	0.903	0.853	0.877
Aroma	1347	0.770	0.682	0.723
ASMOV	1409	0.799	0.772	0.785
BLOOMS	1164	0.954	0.731	0.828
CODI	1023	0.968	0.651	0.779
Ef2Match	1243	0.955	0.781	0.859
GeRMeSMB	528	0.884	0.307	0.456
NBJLM	1327	0.920	0.803	0.858
SOBOM	1246	0.949	0.778	0.855
TaxoMap	1223	0.924	0.743	0.824

Tables 2 and 3 report the SAQ measure results for the various semantic similarity measures listed on the columns. The first number in parentheses after the semantic similarity label indicates the rank of that measure within the row of values for each OA system for only the measures in that table. The second number in parenthesis is the rank of that measure for both tables combined. For the most part each value for a semantic similarity measure had an identical rank across all rows. For example, the lower distance had the highest SAQ value (rank of 1) compared to all other semantic similarity measures across all OA systems in the Table 2. When compared with all measures in both tables, the lower distance was ranked 4th. The Wu-Palmer (WP) measure had the lowest SAQ value (rank of 6) compared to all other semantic similarity measures across all OA systems in Table 2. When compared with all measures in both tables 2 and 3, WP had the lowest SAQ value (rank of 15). If the ranking was not identical across all OA systems, a ranking was only one greater or one less than the most often reoccurring rank in the column. If more than half of one column's ranks had an identical rank value, that rank was used for the SAQ. Note that the first and second rows of the tables are the SAQ results on the partial and full reference alignments provided for the anatomy track.

The SAQ values for the lower distance measure seem to indicate that the alignment quality is extremely good for all these OA systems with it almost being perfect for CODI. There also is very little difference in the OA systems with a range of only

0.00259 between all the values for the SAQ result using the lower cardinality distance measure. The SAQ values for the Wu-Palmer measure seem to indicate that the alignment quality is not as high and has a wider range with a range of 0.01314. But notice all the other SAQ values are greater than 0.90. Another observation is the difference in SAQ results for the two path-based measures. The Leacock-Chodorow agrees more with the IC based results. This experiment indicates there is a substantial difference in SAQ measures depending on what semantic similarity measure is used.

To further investigate this issue, the average WP semantic similarity measure and the average Lin semantic similarity measure was calculated between all 1520*1519/2 pairs of concepts from both the MA and the HA. The WP averages for the MA and HA are 0.015 and 0.074 respectively. The Lin averages for the MA and HA are 0.018 and 0.315 respectively. For the MA, there is little difference in the WP and Lin measures but a substantial difference for the HA. A possible explanation is the MA ontology is not as deep as the HA (maximum depth of 7 vs. 13) and has a less complex structure than the HA (4% vs. 13% of the nodes with multiple parents).

Table 2. SAQ using lower distance, 2 path-based, 3 IC-based semantic similarity measures

OA Systems	**Lower dist 1-D(F) (1) (4)**	**WP (6) (15)**	**LC (4) (13)**	**Lin (3) (12)**	**Resnik (2) (11)**	**JC (5) (14)**
Partial ref	0.99934	0.70647	0.92652	0.94017	0.97274	0.93291
Full ref	0.99902	0.70179	0.92740	0.93633	0.93899	0.92662
AgrMaker	0.99906	0.70234	0.92706	0.93737	0.94009	0.92461
Aroma	0.99718	0.70469	0.92588	0.9385	0.94231	0.91968
ASMOV	0.99866	0.70202	0.92735	0.93836	0.94157	0.92352
BLOOMS	0.99846	0.70301	0.92721	0.94010	0.94296	0.92913
CODI	0.99977	0.70855	0.92660	0.94174	0.94339	0.93684
Ef2Match	0.99936	0.70595	0.92791	0.93816	0.94074	0.92839
GeRMeSMB	0.99936	0.69541	0.92872	0.93268	0.93330	0.92852
NBJLM	0.99907	0.70599	0.92728	0.93797	0.94061	0.92610
SOBOM	0.99921	0.70599	0.92789	0.93988	0.94254	0.93024
TaxoMap	0.99913	0.70816	0.92815	0.93881	0.94154	0.92614

Table 3 shows the results for the SAQ measure using the nine set based semantic similarity measures. The rankings indicate that the downset measures have extremely high SAQ values and the range over all SAQ values using downsets is 0.99996 – 0.99729 = 0.00267. The higher SAQ measure for the downset semantic similarity measures over the upset ones was a surprising result. Intuition suggests that the upset set semantic similarity measures should be better than the downset ones. The rationale is that for downsets, the descendents represent more specific concepts. For example if c is a descendent of a and b, then c inherits features from both A and B but those inherited features may be entirely different and for different purposes. They do not represent common features. But if a and b both have the common ancestor c, then both a and b share c's features.

The unusually high SAQ values for the downset semantic similarity measures which ranked 1st, 2nd and 3rd overall caused further investigations which determined that the downset semantic similarity measures are not useful for the SAQ measure.

When downsets are used in the SAQ, many intersections between the two concepts' sets of descendents are empty. With an empty intersection, all the downset semantic similarity measures produce a 0. This situation is verified by counting the number of cases for the reference alignment where the result for |sim(a,b) – sim(a',b')| is |0 – 0| resulting in a 0 contribution to the $simDiff_{overall}$ total. Close to 50% of the *sumDiff* calculations were |0 – 0|. Not one s*umdiff* produced from any upset semantic similarity measures resulted in a |0 – 0| case. Since the hour set measures include both the upset and the downset, the hour measures also are affected by the extremely large number of cases where there is an empty intersection for the downset. The smaller semantic similarity values then produced smaller *sumdiff* values, thereby, reducing the $simDiff_{average}$. A small $simDiff_{average}$ using the downset measures is not an accurate reflection of the quality of the alignment.

Table 3. SAQ using the nine set-based semantic similarity measures

OA Systems	**Jacc Up (5) (6)**	**Jacc Down (1) (1)**	**Jacc Hr (4) (5)**	**Dice Up (8) (9)**	**Dice Down (2) (2)**	**Dice Hr (6) (7)**	**Inc Up (9) (10)**	**Inc Down (3) (3)**	**Inc Hr (7) (8)**
Partial ref	0.97962	0.99986	0.98311	0.96464	0.99980	0.97030	0.96343	0.99952	0.97962
Full ref	0.97862	0.99982	0.98211	0.96272	0.99974	0.96837	0.96117	0.99922	0.97862
AgrMaker	0.97895	0.99985	0.98243	0.96324	0.99976	0.96896	0.96206	0.99903	0.96363
Aroma	0.97958	0.99978	0.98410	0.96442	0.99966	0.97186	0.96350	0.99729	0.96456
ASMOV	0.97957	0.99971	0.98361	0.96436	0.99961	0.97098	0.96309	0.99870	0.96499
BLOOMS	0.97989	0.99984	0.98359	0.96498	0.99976	0.97105	0.96375	0.99868	0.96507
CODI	0.98020	0.99996	0.98253	0.96551	0.99994	0.96929	0.96384	0.99993	0.96508
Ef2Match	0.97941	0.99987	0.98279	0.96407	0.99980	0.96960	0.96281	0.99948	0.96467
GeRMeSMB	0.97934	0.99990	0.97994	0.96347	0.99985	0.96452	0.96137	0.99932	0.96144
NBJLM	0.97914	0.99984	0.98257	0.96366	0.99975	0.96924	0.96219	0.99908	0.96401
SOBOM	0.97966	0.99986	0.98320	0.96461	0.99978	0.97037	0.96336	0.99926	0.96514
TaxoMap	0.97923	0.99978	0.98284	0.96395	0.99971	0.96982	0.96276	0.99971	0.96463

A question also raised from this experiment is why the lower cardinality distance measure produces the 4th greatest SAQ values. It too uses the concepts' downsets to determine the distance between two concepts. In [30], d_l was chosen with the rationale that the ontologies are more strongly down-branching than up-branching so that down-sets are larger. Siblings deep in the hierarchy are closer together than siblings high in the hierarchy. The intuition behind this seems faulty. The lower cardinality distance suffers from the same problem that the downset set-based measures suffer from – what happens when there is no downset intersection. The SAQ using $|d_l(a,b) - d_l(a',b')|$ translates into the difference between the sum of the number of descendents for a and b and the sum of the number of descendents for a' and b'. Simply because pairs of concepts do not differ greatly in the total number of descendents within their respective ontologies does not mean the mapping is a good mapping. The concepts being mapped could all be leaf or close to leaf nodes but in totally different subtrees of the ontology. Further investigation on the reference alignment shows that the average number of descendents for source and target anchors is 3.2 and 2.8. These averages indicate a very small difference in the number of descendents, and therefore, a very small $simDiff_{average}$.

Table 4. Pearson Correlation with p-value for the SAQ and precision, recall, and f-measure.

SAQ	Precision		Recall		F-measure	
	Corr	p-value	Corr	p-value	Corr	p-value
Lower dist	0.7474974	0.01294	-0.1145901	0.7526	0.02893312	0.9368
WP	0.3600771	0.3068	0.6443854	0.2114	0.7366019	0.01511
LC	0.3710125	0.2912	-0.3711803	0.291	-0.3101822	0.3831
Lin	0.314404	0.3763	0.6277758	0.05198	0.7163154	0.01978
Res	0.131069	0.7182	0.7417546	0.01405	0.7808777	0.007669
JC	0.8058114	0.004886	-0.2055725	0.5688	-0.0002435	0.9995
Jacc Up	0.1846966	0.6095	-0.163976	0.6508	-0.05489762	0.8803
Jacc Down	0.6936438	0.0261	-0.3623896	0.3034	-0.1866459	0.6056
Jacc Hour	-0.2395826	0.505	0.7383087	0.01475	0.6876649	0.02797
Dice Up	0.2132382	0.5542	0.06231594	0.8642	0.171189	0.6363
Dice Down	0.719249	0.01905	-0.404883	0.2458	-0.2190521	0.5432
Dice Hour	-0.2128156	0.555	0.7432621	0.01376	0.6995049	0.02435
Inc Up	0.07231682	0.8426	0.3695583	0.2932	0.4315503	0.213
Inc Down	0.7811095	0.007638	-0.07342486	0.8402	0.08693707	0.8113
Inc Hour	0.1428098	0.6939	0.7285981	0.01685	0.7683169	0.009428

In [30] the Pearson correlation of the lower cardinality distance with the f-measure was given as -0.780. The Pearson correlation for this measure with precision in Table 4 is a 0.7474974. It is positive here since the SAQ is converted into a quality indicator by subtracting from 1. The correlation with f-measure is only 0.02893312. The difference in reported values for the f-measure is unclear unless the reported correlation value in [30] is actually for precision and not f-measure.

From Table 4, all the downset set-based measures had significant correlation with precision and yet, the investigation of the downset measures showed the problem with the huge number of $|0 - 0|$ cases where the downset intersection for both pairs of concepts was empty. The Jiang-Conrath (JC) SAQ is the only other one that had significant correlation with the precision measure. More investigation needs to be done on the SAQ to validate its high correlation with precision.

5 Conclusions and Future Work

This research has investigated the use of semantic similarity measures to evaluate the quality of the mappings produced by OA systems and parallels the work in [30] which experimented with one distance measure between concepts. The research goal is to develop additional means of alignment evaluation that do not depend on a reference alignment. As the experimental results show there are some difficulties with this approach depending on the selected semantic similarity measure.

More research and experiments with SAQ should be undertaken to determine how useful SAQ is for assisting in ontology alignment evaluation especially with respect to precision. SAQ correlation with precision is more intuitive than with recall or f-measure since SAQ is based only on the produced mappings. SAQ has no knowledge of missed mappings.

Further investigation is needed to determine how much poor mappings affect the resulting SAQ and identify and eliminate these very poor mappings if the semantic similarity difference is above a specified threshold. The semantic similarity difference operation might be more useful in the alignment process itself than in the evaluation of the final mappings. The OA systems in this experiment specifically use semantic similarity with a knowledge source and not between concepts in the source and target ontologies in their matching algorithms. If a semantic similarity measure is used in an OA system's matching process, research is needed to see how the SAQ evaluation of its mapping result may be biased based on the selected measure.

Acknowledgments. The authors acknowledge the implementation of the SAQ measure was embedded in AgreementMaker and Cosmin Stroe served as a consultant.

References

1. Giunchiglia, F. Shvaiko, P. and Yatskevich, M**.:** S-Match: an algorithm and an implementation of semantic matching**.** Technical Report DIT-04-015, Department of Information Engineering and Computer Science, University of Trento. Proc. of the first European Semantic Web Symposium (ESWS) (2004)
2. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Ontology Matching: A Machine Learning Approach. In: Handbook on Ontologies in Information Systems. pp. 397—416. Springer (2003)
3. Rodriguez, M. A., Egenhofer, M. J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering, vol. 15, issue 2, pp, 442--456 (2003)
4. Tversky, A.: Features of Similarity. Psychological Rev.*,* 84, pp. 327--352 (1977)
5. Sabou, M., d'Aquin, M., Motta, E.: Exploring the Semantic Web as Background Knowledge for Ontology Matching. J. Data Semantics 11, pp. 156--190 (2008)
6. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J.: Introduction to Wordnet: An on-line Lexical Database. International Journal of Lexicography 3(4), pp. 235—244 (1990)
7. Euzenat, J., Valtchev, P.: An integrative proximity measure for ontology alignment. In: Proc. ISWC-2003 Workshop on semantic information integration, Sanibel Island (FL US), pp. 33--38 (2003)
8. Tang, J., Liang, B.Y., Li, Juanzi, Wang, Kehong: Risk Minimization based Ontology Mapping. 2004 Advanced Workshop on Content Computing (AWCC). LNCS, vol. 3309, pp. 469--480. Springer-Verlag (2004)
9. Li, Juanzi, Tang, Jie, Yi, Li, Luo, Qiong: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Transactions on Knowledge and Data Engineering, vol. 21 issue. 8. pp. 1218--1232 (2009)
10. Unified Medical Language System (UMLS) http://umlsks.nlm.nih.gov
11. Cruz, I. F., Sunna, W.: Structural Alignment Methods with Applications to Geospatial Ontologies. Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications vol. 12 no. 6, pp. 683—711 (2008)
12. Cruz, I F., Palandri Antonelli, F., Stroe, C.: AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. PVLDB,vol. 2, no. 2, pp. 1586--1589 (2009).
13. Cruz, I. F., Stroe, C., Caci, M., Caimi, F., Palmonari, M., Palandri Antonelli, F., Keles, U. C.: Using AgreementMaker to Align Ontologies for OAEI. In: ISWC International

Workshop on Ontology Matching (OM), ser. CEUR Workshop Proceedings vol. 689, pp. 118—125 (2010)
14. Jean-Mary, Y.R., Shironoshita, E. P., Kabuka, M. R.: Ontology matching with semantic verification. Web Semantics, vol 7 issue 3. pp. 235--251 (2009)
15. Posse, C., Sanfilippo, A., Gopalan, B., Riensche, R., Beagley, N., Baddeley, B.: Cross-Ontological Analytics: Combining Associative and Hierarchical Relations in the Gene Ontologies to Assess Gene Product Similarity. International Conference on Computational Science (2), pp. 871—878 (2006)
16. d'Amato , C., Fanizzi, N., Esposito, F.: A dissimilarity measure for ALC concept descriptions. In: Proc. ACM Symposium on Applied Computing (SAC), ACM. pp. 1695–1699 (2006)
17. Araujo, R., and Pinto, H. S. Towards semantics-based ontology similarity. In: Proc. Workshop on Ontology Matching (OM), International Semantic Web Conference (ISWC). (2007)
18. Janowicz, K., Wilkes, M.: SIM-DL_A: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-Concept to Inter-Instance Similarity. In: The 6th Annual European Semantic Web Conference (ESWC2009). Lecture Notes in Computer Science 5554, Springer. pp. 353-367 (2009)
19. Cross, V. and Yu, Xinran: Investigating Ontological Similarity Theoretically with Fuzzy Set Theory, Information Content, and Tversky Similarity and Empirically with the Gene Ontology. In: Proc. of the 5th International Conference on Scalable Uncertainty Management, Dayton OH (2011)
20. Rada R, Mili H, Bicknell E, Blettner M: Development and Application of a Metric on Semantic Nets. In: IEEE Transaction on Systems, Man, and Cybernetics vol. 19, pp. 17–-30 (1989)
21. Leacock C. and Chodorow, M.: Combining local context and WordNet Similarity for Word Sense Identification. In: WordNet: An Electronic Lexical Database. Fellbaum, Ed. Cambridge, MA: MIT Press, pp. 265--283 (1998)
22. Su, Xiaomeng: Semantic Enrichment for Ontology Mapping, Ph.D. Thesis, Dept. of Computer and Information Science, Norwegian University of Science and Technology (2004)
23. Wu Z, Palmer M. S.: Verb Semantics and Lexical Selection. In: Proc. of the 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133--138. (1994)
24. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in Taxonomy. In: Proc. of the 14th International Joint Conference on Artificial Intelligence, pp, 448--453 (1995)
25. Seco N, Veale T, Hayes J.: An Intrinsic Information Content Metric for Semantic Similarity in Wordnet. In: ECAI. pp. 1089--1090 (2004)
26. Lin D.: An Information-theoretic Definition of Similarity. In: Proc. of the 15th International Conference on Machine Learning. Morgan Kaufmann. pp. 296--304 (1998).
27. Giuseppe, P., Talia D.: UFOme: An Ontology Mapping System with Strategy Prediction Capabilities. Data Knowl.Eng. vol. 69 no. 5, pp. 444--71 (2010)
28. Jiang J, Conrath D: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Proc. of the 10th International Conference on Research on Computational Linguistics, Taiwan (1997)
29. Cross V., Sudkamp, T.: Similarity and Compatibility in Fuzzy Set Theory, Heildelberg: Physical-Verlag (2002)
30. Joslyn, Cliff A., Paulson, P., White, A.: Measuring the Structural Preservation of Semantic Hierarchy Alignment. In: ISWC International Workshop on Ontology Matching. CEUR-WS. (2009).
31. Euzenat, J.et al.: The Results of the Ontology Alignment Evaluation Initiative 2010. Ontology Matching Workshop, International Semantic Web Conference (2010)