Fuzzy Set and Semantic Similarity in Ontology Alignment

Valerie Cross
Computer Science and Software Engineering
Miami University
Oxford, OH USA
crossv@muohio.edu

Xueheng Hu
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN USA
xhu2@nd.edu

Abstract – **A challenge for the Semantic Web is enabling information interoperability between related but heterogeneous ontologies. Ontology alignment (OA) addresses this challenge by identifying correspondences between entities in different ontologies. The traditional OA evaluation strategy uses a gold standard reference alignment created by a domain expert. The problem is that often a reference alignment may not exist. The current use of semantic similarity measures in the OA process and proposals for their use in the OA evaluation task are presented. Many semantic similarity measures are derivative of fuzzy set similarity measures. A general semantic alignment quality (SAQ) measure is developed and used on the alignment results of 10 different OA systems produced on the anatomy track of the 2010 OA evaluation initiative. The SAQ results indicate much variation in performance depending on the selected semantic similarity measure. Problems with using several semantic similarity measures in SAQ are further investigated and the findings are discussed.**

Keywords-ontology alignment; semantic similarity; fuzzy set similarity; ontology alignment evaluation initiative (OAEI)

I. INTRODUCTION

Although numerous disciplines such as psychology, biology, and statistics have been using a variety of similarity measures between objects, the advent of the Semantic Web and particularly bioinformatics research has renewed interest in measuring similarity between concepts defined in an ontology. In this context, similarity has been referred to as semantic similarity or also as ontological similarity. An important challenge facing the Semantic Web is facilitating exchange of information among related but heterogeneous ontologies, that is, information interoperability. A key means of dealing with this challenge is the ontology alignment (OA) process. A primary operation in this process is similarity assessment between entities. A wide variety of similarity measures have been used, but semantic similarity measures have not been capitalized on in OA. This paper investigates the use of semantic similarity measures in this process and their application in a proposed semantic alignment quality (SAQ) measure to evaluate the alignment results of OA systems. The SAQ measure does not require a reference alignment as does the standard OA performance measures of precision, recall and the f-measure [1].

The paper begins with Section II briefly describing the ontology alignment process and the importance of feature selection to assess similarity between entities in the two ontologies being aligned. Section III presents more details on similarity measures in the OA process. It also shows how fuzzy set similarity measures are the basis for many of the semantic similarity measures. Section IV presents the proposed method of using semantic similarity measures in the SAQ measure to evaluate the alignment results produced by an OA system. The implementation of the SAQ measure is described in section V. Section VI discusses the experiments and their results using the SAQ measure on the alignments produced by ten different OA systems that participated in the anatomy track of the 2010 ontology alignment evaluation initiative (OAEI) [2]. Section VII concludes by briefly summarizing the research and outlining future work to use semantic similarity measures in the OA process.

II. ONTOLOGY ALIGNMENT OVERVIEW

The term ontology alignment as used in this research and the format of the output of the ontology alignment process is based on the description in [1]. Ontology alignment takes as input a source ontology O_s and finds for each of its entities e_s (concept, relation, or instance) an entity e_t in the target ontology O_t that is closely related to it or has the same meaning. Typically, when only looking for entities with the same meaning, the alignment process simply determines a one-to-one equality or identity relation that exists between a source entity and a target entity. Other kinds of relations besides equality such as subsumption may also be determined between two entities.

A general ontology alignment function requires a set of entities, E, the set of possible ontologies O, and the set of possible alignment relations, M. It is a partial function: $genalign: E \times O \times O \rightarrow E \times M$. The result of the alignment process from most ontology alignment systems is a set of correspondences (e_s, e_t, m, l) such that the relation m holds

between e_s and e_t with a confidence level l. The confidence level is optional and is in (0, 1].

Alignment results from most OA systems are uniformly produced using the Alignment API [3]. This API implemented in Java allows 1) storing, finding, and sharing alignments, 2) piping alignment results through algorithms to improve an existing alignment, 3) manipulating alignments such as thresholding and hardening on confidence levels, 4) generating processing output such as transformations, axioms, and rules, and 5) comparing alignments.

The OA process followed by the typical OA system consists of several steps [1]. The first step is selecting the features of the entities in the ontology definition that are to be used in determining the similarity between entities of the two ontologies. The features of an entity from the source ontology are compared to those of an entity in the target ontology. When comparing entities, an OA system might only compare entities of the same type, for example concepts with only concepts or may compare entities of different types with each other. Most only compare entities of the same type.

The feature selection step of the OA process is critical. As Goodman [4] points out, assessing similarity between two objects *x* and *y* is vague and meaningless since the "'is similar to' functions as little more than a blank to be filled." Asking the question "'How similar are *x* and *y*?' begs the answer to a subtly different question 'How are *x* and *y* similar?'" [5]. For OA systems, the answer to this question takes many forms but is typically implemented in various matchers of an OA system. Each OA matcher produces its own similarity value for a candidate alignment pair which depends on the feature(s) selected in answering the question 'How are *x* and *y* similar?'.

After the various similarity values have been calculated, the overall similarity for an entity pair is produced using an aggregation function of the individual similarity values. The overall similarity values for entity pairs are then used to derive the alignments for the two ontologies. Iteration may occur in the alignment process. For each iteration, the similarities of a candidate alignment are recalculated based on the similarities of neighboring entity pairs. The iteration terminates when no new alignments are proposed. Not all steps may be necessary in the iteration. For example, some feature similarities might only be determined in the first iteration and not influenced by similarity between neighboring entity pairs.

From the above description of the OA process, the two primary operations used are feature similarity calculation between two entities followed by aggregation of the individual feature similarity values. Iteration of the alignment process is necessary if the calculation of a similarity value for a feature is influenced by other previously determined feature similarity values. Similarity calculation and then overall aggregation of similarity values are also two of the primary operations in fuzzy approximate reasoning. Much overlap exists in the kinds of similarity measures and aggregation operators used in both ontology alignment and fuzzy approximate reasoning [6].

Early OA work focused simply on one feature of an entity: its string name. Different string similarity measures such as prefix, suffix, edit distances and n-gram have been used to determine similarity between the string labels of entities. But creation of ontologies even in the same domain still allows for wide variance in the terminology. GLUE [7] was one of the first OA systems to combine several different learners (similar to what is currently being called a matcher) to establish mappings between entities. It combines a learner using instance information with another learner that uses a concept's complete list of ancestor concepts from the ontology root to the concept itself to determine similarity between concepts in two different ontologies. Now, however, a wide variety of matchers are used in OA systems. These matchers belong to various categories depending on the context of the similarity measurement, such as lexical, structural, or extensional matchers [8].

Many of these matchers have as their underpinnings the various features of entities such as a concept's attributes or its neighboring concepts that could be used for "filling in the blank" in a similarity function. The similarity function proposed in [9] is a slight variation of Tversky's parameterized ratio model of similarity [10]. These kinds of similarity measures have also been utilized in determining overall ontology similarity, the global similarity between two ontologies [11]. Such similarity measures have been adapted for use in matchers of various OA systems.

III. Fuzzy Set and Semantic Similarity in OA

Semantic similarity differs from the similarity measures typically used within OA since it assesses the similarity between two concepts within a single ontology instead of across two ontologies. Early on simple path distance, i.e., the count of the number of edges or nodes, between two concepts within the graph structure of the ontology was proposed for measuring the semantic similarity of the two concepts [12]. The edges typically represent the subsumption relation or the part-of relation between concepts. Early measures assigned a uniform weight of 1 to all the edges regardless of an edge's position in the ontology graph. The simple path-based distance measure [12] was converted into a similarity measure by Leacock and Chodorow [13]. They normalized the path length between the two nodes by dividing by twice the maximum depth of the ontology and then taking the negative logarithm of this value. Wu and Palmer [14] improved upon these early path-based similarity measures by incorporating the distance of each concept from the root concept.

Another approach to semantic similarity is based on using a measure of information content (IC) for a concept. IC measures how specific a concept is within a given ontology and assumes a well defined complete as possible ontology. The more specific a concept is the higher its information content. The more general a concept is the lower its information content. IC has been determined by either a corpus-based or an ontology-based method. The corpus-based IC uses an external resource such as an associated corpus for

the problem domain. The corpus-based IC measure for concept c [15] is given as

$$IC_{corpus}(c) = -\log p(c) \quad (1)$$

The value $p(c)$, the probability of concept c, is determined by the frequency count of the concept, i.e. the count of the number of its occurrence within the corpus. The count is the number of occurrences in the corpus of all words representing that concept. The frequency of the concept also includes the total frequencies of all its children concepts.

The ontology-based IC method simply uses the structure of ontology itself to determine a concept's IC value. The ontology-based IC [16] for concept c is defined as

$$IC_{ont}(c) = 1 - \frac{\log(num\_desc(c)+1)}{\log(max_{ont})} \quad (2)$$

where $num\_desc(c)$ is the number of descendants for concept c and max_{ont} is the maximum number of concepts in the ontology. It is normalized in [0...1] with the maximum of 1 for all the leaf concepts and deceases to the minimum of 0 at the root concepts.

The first IC-based semantic similarity measure is defined as the maximum information content two concepts share [15]. The common ancestor of the two concepts having the maximum IC value must be found and its IC value is taken as the semantic similarity between the two concepts. An improvement to Resnik's shared IC approach was proposed by Lin [17]. It uses not only the maximum shared information content between the two concepts but also each concept's individual information content.

Such semantic similarity measures are currently being used within a background knowledge source, i.e., a reference or mediating ontology such as WordNet and the Unified Medical Language System (UMLS) to aid in finding a mapping between concepts in two different ontologies. For example, OLA [18] uses a modified version of the Wu-Palmer semantic similarity measure with WordNet to determine lexical similarity between a pair of identifiers. For a pair of concepts, iMapper [19] finds each one's descriptive label in WordNet and uses a simple path based semantic distance between the WordNet concepts. The similarity value between the two concepts in the different ontologies being aligned may then be strengthened based on the distance of the two concepts in WordNet. In [20], semantic similarity measure is used in the filtering of mappings. The techniques are adapted from the PowerMap WordNet based algorithm [21] with the Wu-Palmer measure being used. ASMOV [22] uses both WordNet and UMLS as reference ontologies and the Lin semantic similarity measure to determine the lexical similarity between concept labels. The WordNet matcher of UFOme [23] also locates the concept labels in WordNet and employs the Lin semantic similarity measure.

These path-based and information content semantic similarity measures incorporate the ontological structure and specificity of the concept into the similarity measurement. But as pointed out in [9] another approach views an entity as being described by a set of features. Similarity between concepts can then be measured using the set-based approach presented by Tversky's parameterized ratio model [10] given as

$$S_{Tverksy\text{-}ratio}(X, Y) = \frac{f(X \cap Y)}{f(X \cap Y) + \alpha f(X - Y) + \beta f(Y - X)}. \quad (3)$$

In the model, X and Y represent sets describing features of two objects x and y. $(X \cap Y)$ represents the common features that describe both x and y. $(X - Y)$ represents the features describing only object x. $(Y - X)$ represents the features describing only object y. The value $f(X)$ for object x is considered a measure of the overall salience of that object. Factors adding to an object's salience include "intensity, frequency, familiarity, good form, and informational content" [10]. The function f is an additive function on disjoint sets, i.e., whenever X and Y are disjoint sets and when all three terms are defined, then $f(X \cup Y) = f(X) + f(Y)$. Set cardinality is such a function f.

Parameters α and β may be used to select which object is the reference object, i.e., the one to which similarity is being determined. If x is the reference object, then α should be greater than β since the features describing x but not y are consider more significant than those describing y but not x. With $\alpha = \beta = 1$, $S_{Tversky}$ becomes the Jaccard index given as

$$S_{jaccard}(X,Y) = \frac{f(X \cap Y)}{f(X \cup Y)}. \quad (4)$$

With $\alpha = \beta = ½$, $S_{Tverksy\text{-}ratio}$ becomes Dice's coefficient of similarity defined as

$$S_{dice}(X, Y) = \frac{2 \times f(X \cap Y)}{f(X) + f(Y)}. \quad (5)$$

With $\alpha = 1$, $\beta = 0$, $S_{Tverksy\text{-}ratio}$ becomes the degree of inclusion for X, that is, the proportion of X overlapping with Y

$$S_{inclusion}(X, Y) = \frac{f(X \cap Y)}{f(X)}. \quad (6)$$

The relationships between the path-based and IC-based semantic similarity measures to feature or set-based similarity measures such as the Tversky models [10] or fuzzy set compatibility measures [24] were initially not examined by researchers. These historical semantic similarity measures are shown to be variations of feature-based similarity measures when one examines how the "blanks" are filled in, i.e., how the features are selected to describe a concept in ontology. A set of features for a concept may be given a fuzzy set-theoretic interpretation, and as such, many of the early semantic similarity measures can be seen as fuzzy set similarity measures [25]. Here, briefly this interpretation is summarized since several of the semantic similarity measures in the experimental results reported in section 6 are based on this interpretation and are basically fuzzy set similarity measures.

Each concept c in an ontology can be described by a fuzzy set. Which fuzzy set is used depends on how one proceeds to select the concept's features, for example, its properties, its children, its parents, etc [9]. Another important aspect to the

fuzzy set is the method used to determine membership degree of its elements. For example, one fuzzy set describing the concept *c* is based on selecting its ancestor set with the membership degree as a function of the IC associated with each concept, that is,

$$F_{anc+(c)}(c_j) = \{IC(c_j)/c_j \mid c_j \text{ is an ancestor of } c \text{ or } c \text{ itself}\} \quad (7)$$

where the + indicates to include the concept *c* itself in the set. Here the function on IC is simply the identify function. This fuzzy set specifies each element c_j and its respective membership $IC(c_j)$ in the fuzzy set. A concept can be described using many different sets; for example, instead of the ancestor set, the descendent set could be used or the set of links for a path associated with concept *c* within the ontology.

Given a fuzzy set interpretation to describe a concept, the Jaccard fuzzy set compatibility measure can be used to calculate IC-based semantic similarity between two concepts represented by their fuzzy sets of ancestors $F_{anc+(c1)}$ and $F_{anc+(c_2)}$. The set intersection uses a t-norm, typically minimum, and the set union uses a t-co-norm, typically maximum. The Jaccard semantic similarity measure with the anc+ set selected then becomes

$$sim_{JacAnc}(c1, c2) = \frac{\sum_{c \in F_{anc+(c1)} \cap F_{anc+(c2)}} IC(c)}{\sum_{c \in F_{anc+(c1)} \cup F_{anc+(c2)}} IC(c)} \quad (8)$$

To further explain, the min and max fuzzy set operators do not need to be explicitly used for the intersection and union operators since IC as typically calculated for a concept is the same in each fuzzy set because it is based on a function of the concept's number of descendents. If another membership function is used and a concept's membership could differ in the two fuzzy sets, then the min and max operators would be needed.

IV. Using Similarity in OA Evaluation

Although semantic similarity measures have been used in some OA systems with reference ontology to assist in determining lexical similarity between concepts, they have not been used in the evaluation of alignment results. The standard method for the evaluation of OA results requires a gold standard reference alignment which has typically been created by a set of human domain experts. It is considered to be a correct and complete set of mappings between the two ontologies. The evaluation of the OA result is then determined using three standard information retrieval measures adapted for OA. These are: precision, recall, and f-measure [1]. The precision is the ratio of the number of correct mappings generated by the OA system and the total number of mappings generated by the OA system. The recall is the ratio of the total number of correct mappings generated by the OA system and the total number of mappings in the reference alignment. The f-measure is a parameterized combination of the former two.

This paper reports the results of experiments to use semantic similarity measures for OA evaluation purposes as an alternative or in addition to the standard OA evaluation measures. The semantic alignment quality (SAQ) measure is to determine for the source ontology O_s and the target ontology O_t how well each pair of mappings, (s_i, t_i) and (s_j, t_j) maintains the same semantic similarity between the corresponding concepts pairs (s_i, s_j) in the source and (t_i, t_j) in the target, i.e., $|sim(s_i,s_j) - sim(t_i,t_j)|$ should be close to 0 for all aligned pairs. These individual differences contribute to a total overall SAQ measure.

More precisely, assume that the set of mappings M is {(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

$$simDiff(m_i, m_j) = | sim(s_i, s_j) - sim(t_i, t_j) | \quad (9)$$

with s_i and s_j being source entities and t_i and t_j being target entities 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 pairs and target pairs is calculated as

$$simDiff_{overall}(M) = \sum_{m_i, m_j \in M} simDiff(m_i, m_j). \quad (10)$$

The average difference of semantic similarity is calculated over all (m_i, m_j) pairs in M where $i \neq j$ as

$$simDiff_{average}(M) = \frac{simDiff_{overall}(M)}{C_2^N} \quad \text{where } N = |M|. \quad (11)$$

The SAQ measure is $1 - simDiff_{average}$ so that the smaller the average of the semantic similarity differences, the closer the SAQ is to 1. A high SAQ value is expected to indicate the alignment pairs (s_j, t_j) have been well chosen. Notice though that the SAQ is generically defined in that any semantic similarity measures may be substituted for the *sim* function.

This generic approach using semantic similarity for ontology alignment evaluation provides more flexibility than that in [26] where a distance measure between concepts within a concept lattice is proposed. There an ontology is viewed as a concept lattice and a distance between two concepts *a* and *b* is defined based on a selected set of related concepts as

$$d_{set}(a,b) = (|set(a)| + 1) + (|set(b)| + 1) - 2*\max_{latticeOp\text{-}c}[(set(c)+1] \quad (12)$$

where *set* can be either the ancestor or the descendent set of the concept. The lattice operator *latticeOp* is determined by the selected set. When the set of ancestors for a concept is used, the lattice operator is the join. The *c* concept must be the join concept between *a* and *b* with the maximum number of ancestors. The distance is referred to as the upper cardinality distance. When the set is the descendents for a concept, the lattice operator is the meet. The *c* concept must be the meet concept between *a* and *b* with the maximum number of descendents. This distance is then referred to as the lower cardinality distance. In [26] only the lower cardinality distance is used in an experiment to determine the performance of this measure for evaluating ontology alignment results.

The work presented here in this paper generalizes that in [26] which is based on a concept lattice view of an ontology.

The ontology research community is more accustomed to working with semantic similarity measures in ontology research than the newly proposed distance measures based on ordered lattice theory. This paper examines the use of a wide variety of semantic similarity measures for the purpose of evaluating ontology alignment results and also includes the proposed lower cardinality distance measure.

V. Semantic Alignment Quality Architecture

The major input to the SAQ system besides user settable parameters is the alignment result of an OA system and the two ontologies being aligned. Many existing OA systems process their own alignment results in order to calculate the standard performance measures. Rather than re-implement this needed functionality, a decision was made to investigate existing OA systems to determine into which system to embed the SAQ measure. Based on examining several OA systems and contacting their developers, AgreementMaker [27] was selected. AgreementMaker has the following modules that facilitate the SAQ implementation: 1) an ontology loading module that loads and parses the source and the target ontologies, 2) an alignment loading module that loads existing alignment result files in standard Alignment API format and parses the alignment results to obtain the mapping information, and 3) an extensible evaluation module.

The GUI of AgreementMaker was also modified as seen in Figure 1 to provide a drop-down list for the user to select the SAQ metric to use in the evaluation of the OA systems.

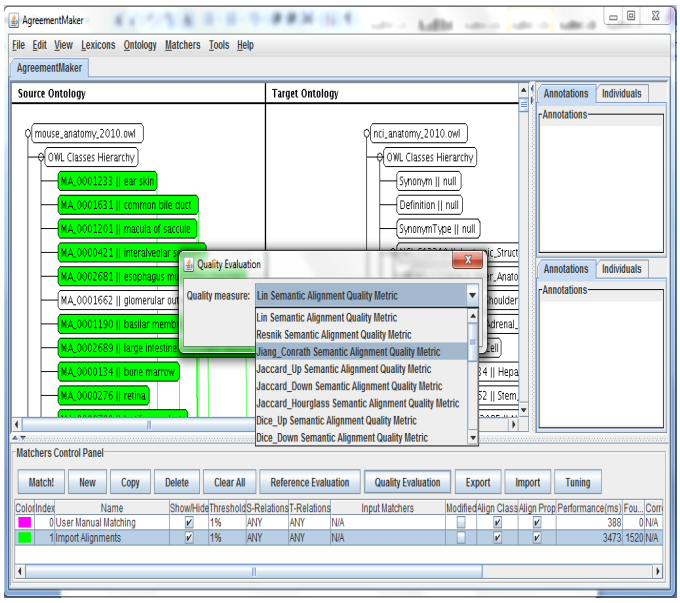

Figure 1. SAQ User Interface within AgreementMaker

Each SAQ metric is implemented as a concrete instance of the AgreementMaker API: AbstractQualityMetric. The AgreementMaker class: QualityMetricRegistry is then updated with each SAQ metric. The SAQ system has 14 semantic similarity measures: the information Content measures of Lin [17], Resnik [15], and Jiang & Conrath [28], the set-based measures of Jaccard, Dice, and inclusion [10], and the path-based measures of Wu & Palmer [14] and Leacock & Chodorow [13]. There are a total of nine different set-based measures since each of the three set-based measures has three variations depending on which set is selected for describing a concept. The three sets used are the ancestor set (up), the descendent set (down) and the hourglass set which is the union of the ancestor and the descendent sets. In addition, the lower cardinality distance measure is included.

VI. Experiments Using the OAEI Anatomy Track

Current OA systems typically compete in the ontology alignment evaluation initiative (OAEI) which has been conducted annually since 2004 [2]. This competition uses the same set of test cases that are over four different tracks: anatomy, benchmarks, conference, and directory. This consistent testing methodology allows evaluation of these systems so that interested users can compare their performance and developers can improve their OA systems based on these evaluations.

The anatomy track is used in the experiments with the SAQ measure. This track was selected because it uses two real-world ontologies from the biomedical domain, the NCIT human anatomy (HA) ontology and the mouse anatomy ontology (MA) which are significantly larger than the ontologies in the other tracks. The MA ontology has 2744 classes while the HA ontology has 3304 classes. This track produces many more mappings in the ontology alignment results than those of the other tracks. Since the gold-standard reference alignment between HA and MA is published and the precision, recall and f-measure of each OA system participating in the OAEI 2010 anatomy track are also available, the SAQ measures for each OA system alignment result may be compared to the standard OA performance measures used in the OAEI.

The anatomy track has 4 subtasks. The subtasks are to generate alignments that 1) are as good as possible with respect to the f-measure, 2) favour precision over recall, 3) favour recall over precision, and 4) use a partial reference alignment as additional input to the alignment process. Since only subtask 1 is obligatory for all OAEI participants, it is used in this reported SAQ experiment in order to have the greatest number of OA system results for comparison purposes.

Table 1 lists the OA systems in alphabetical order along with the number of mappings they produced and their performance measures on the first subtask of the anatomy track of OAEI 2010. The rank number following each OA system indicates its rank with respect to its f-measure score. AgreementMaker performs the best on subtask 1 with three other systems Ef2Match, NBJLM and SOBOM clustered around a 0.860 value. Although subtask 1 favours the f-measure, all OA systems except for Aroma, ASMOV, and GeRMeSMB produce precision values over 0.900, and all OA systems except for Aroma, CODI, and GeRMeSMB produce recall values over 0.700. Note that CODI produced the greatest precision value overall but was a poor performer with respect to recall.

TABLE 1. OA SYSTEMS OAEI 2010 ANATOMY TRACK PRECISION (P), RECALL (R), F-MEASURE (F)

OA Systems	Rank	# of Mappings	P	R	F
AgrMaker	1	1436	0.903	0.853	0.877
Aroma	9	1347	0.770	0.682	0.723
ASMOV	7	1409	0.799	0.772	0.785
BLOOMS	5	1164	0.954	0.731	0.828
CODI	8	1023	0.968	0.651	0.779
Ef2Match	2	1243	0.955	0.781	0.859
GeRMeSMB	10	528	0.884	0.307	0.456
NBJLM	3	1327	0.920	0.803	0.858
SOBOM	4	1246	0.949	0.778	0.855
TaxoMap	6	1223	0.924	0.743	0.824
Average		1194.6	0.903	0.710	0.784

The purpose of the experiments with the SAQ metric on subtask 1 is to investigate the difference in performance of the various semantic similarity measures in the SAQ metric and examine whether the notion of SAQ corresponds with the standard performance measures used to evaluate OA results.

The results for all participating OA systems using the SAQ metric with each of the 14 semantic similarity measures implemented within Agreementmaker's quality evaluation module are shown in Table 2, Table 3a and Table 3b with results rounded to three decimal positions. Table 2 contains the path-based and information content semantic similarity measures. Included as column 1 in Table 2 are the results using the concept lattice lower cardinality distance measure 1-D(F) [26]. Table 3a contains the Up (ancestors) and Down (descendents) fuzzy set-based semantic similarity measures. Table 3b contain the Hour (union of ancestors and descendents) fuzzy set-based semantic similarity measures.

Below the column label of the SAQ measure are two numbers. The first number indicates the rank of the SAQ score within the row of values for each OA system for only the measures in that table (where Table 3a and Table 3b are considered one table). The second number in parenthesis specifies the rank of that measure over all 15 reported measures. A rank of 1 indicates the OA system produced the greatest SAQ measure.

TABLE 2. SAQ FOR LOWER CARDINALITY, PATH-BASED AND INFORMATION-CONTENT MEASURES

OA Systems	1-D(F) (1) (3.5)	WP (6) (15)	LC (4) (13)	Lin (3) (12)	Resnk (2) (11)	JC (5) (14)
AgrMaker	0.999	0.702	0.927	0.937	0.940	0.925
Aroma	0.997	0.705	0.926	0.939	0.942	0.920
ASMOV	0.999	0.702	0.927	0.938	0.942	0.924
BLOOMS	0.998	0.703	0.927	0.940	0.943	0.929
CODI	1.000	0.709	0.927	0.942	0.943	0.937
Ef2Match	0.999	0.706	0.928	0.938	0.941	0.928
GeRMeSMB	0.999	0.695	0.929	0.933	0.933	0.929
NBJLM	0.999	0.706	0.927	0.938	0.941	0.926
SOBOM	0.999	0.706	0.928	0.940	0.943	0.930
TaxoMap	0.999	0.708	0.928	0.939	0.942	0.926

An observation in most cases is the rank for a measure across all OA systems is consistent. For example, the lower distance measure ranked first in Table 2 with the highest SAQ value compared with the path-based and IC-based semantic similarity measures for all OA systems. When the two tables are combined, there is a tie between JaccD and DiceD since the results are identical and so they have a rank 1.5 over all measures. IncD and 1-D(F) are almost nearly identical with 0.001 higher values for IncD for the BLOOMS and TaxoMap systems so they are considered as tied and ranked 3.5 over all measures.

TABLE 3a. SAQ WITH UP AND DOWN SET SIMILARITIES

OA Systems	JaccU (5) (6)	DiceU (8) (9)	IncU (9) (10)	JaccD (1.5) (1.5)	DiceD (1.5) (1.5)	IncD (3) (3.5)
AgrMaker	0.979	0.963	0.962	1	1	0.999
Aroma	0.98	0.964	0.964	1	1	0.997
ASMOV	0.98	0.964	0.963	1	1	0.999
BLOOMS	0.98	0.965	0.964	1	1	0.999
CODI	0.98	0.966	0.964	1	1	1
Ef2Match	0.979	0.964	0.963	1	1	0.999
GeRMeSMB	0.979	0.963	0.961	1	1	0.999
NBJLM	0.979	0.964	0.962	1	1	0.999
SOBOM	0.98	0.965	0.963	1	1	0.999
TaxoMap	0.979	0.964	0.963	1	1	1

TABLE 3b. SAQ WITH HOURGLASS SET SIMILARITIES

OA Systems	JaccH (4) (5)	DiceH (6) (7)	IncH (7) (8)
AgrMaker	0.982	0.969	0.964
Aroma	0.984	0.972	0.965
ASMOV	0.984	0.971	0.965
BLOOMS	0.984	0.971	0.965
CODI	0.983	0.969	0.965
Ef2Match	0.983	0.970	0.965
GeRMeSMB	0.980	0.965	0.961
NBJLM	0.983	0.969	0.964
SOBOM	0.983	0.970	0.965
TaxoMap	0.983	0.970	0.965

Notice that all the down sets measures and the lower cardinality distance produce extremely high if not perfect SAQ values. Based on the intuition that small differences, i.e., $|sim(s_i,s_j) - sim(t_i,t_j)|$ indicate each pair of mappings, (s_i, t_i) and (s_j, t_j) maintains the same semantic similarity between the corresponding concept pairs (s_i, s_j) in the source and (t_i, t_j) in the target, then the alignment quality is extremely high for all OA systems. Essentially, using the downset (descendents) semantic similarity measures, one cannot tell a difference in performance between the OA systems.

The extremely high and perfect 1.0 results produced by the downset semantic similarity measures and the lower cardinality distance in the SAQ for all the OA systems seems

implausible considering that the f-measure shows a difference in the performance of GeRMeSMB and AgreementMaker systems. Further investigating the downset semantic similarity measures determined these measures are not good at measuring the similarity of concepts, at least in the OAEI anatomy track. This result agrees with intuition since for downsets, descendents represent more specific concepts. If *c* is a descendent of both *a* and *b*, then *c* inherits features from both *a* and *b*. The problem is that those inherited features may be entirely different and for different purposes because they are coming from two different concepts. Having a common descendent requires intentionally creating a concept that represents a combining of two or other concepts and is not as natural as finding a common ancestor for two concepts.

Investigating the downset measures revealed the empty intersection problem. For the MA and HA ontologies, an empty intersection between two concepts' descendant sets is a fairly common. An empty intersection, produces for all downset semantic similarity measures a 0. To verify this, the full reference alignment between the MA and HA was processed and the number of cases where $|sim(a, b) - sim(a', b')|$ is actually $|0 - 0|$ was counted. After the tally, it was discovered that they make up nearly 50% of the *simDiff* calculations and thus they contribute a 0 to the $simDiff_{overall}$. These cases greatly reduce the $simDiff_{average}$. According to the intuition behind the SAQ measure a low $simDiff_{average}$ value, i.e., a high SAQ value, indicates that the alignment is of high quality. For the downset measures, however, the empty intersection problem reveals that the high SAQ values are not a valid indication of high alignment quality.

Lower cardinality distance measure also uses the downsets of two concepts to determine the distance between those two concepts. In [26] the downset is chosen with the explanation that ontologies are more strongly down-branching than up-branching. They reasoned that siblings deep in the ontology are closer together than siblings that are high in the ontology. Semantic similarity research very early on recognized that the distance between siblings lower in the ontology should be smaller than that between siblings higher up. Selecting the lower cardinality based on this fact is not justifiable. Using a concept's ancestors in the measurement of concept similarity is more intuitive due to concepts sharing common features when they share a common ancestor.

Using the lower cardinality distance also can have the situation where the two concepts *a* and *b* do not have a meet concept *c* except for the artificial bottom concept of the lattice. When this occurs, the lower cardinality distance between concepts *a* and *b* basically becomes the sum of the number of descendents of *a* and *b* and also adding one for each concept. When both *a* and *b* are leaf concepts and siblings, they have the same distance as when both are leaf concepts and cousins, i.e., children of sibling concepts. Then *simDiff* for the aligned concept pairs (s_i, t_i) and (s_j, t_j) becomes the difference between the total number of descendents for s_i, and s_j and the total number of descendants for t_i and t_j.

To validate these notions, the reference alignment was again used to determine the average numbers of descendants for the source and target anchors, only 3.2 and 2.8 respectively. The low averages indicate a very small overall difference in the sum of the number of descendants and, therefore, results in a very small $simDiff_{average}$ and a high SAQ value. The result explains why the smallest lower cardinality distance SAQ measure is only 0.997, that is, nearly perfect. The SAQ with lower cardinality distance is not a good measure of alignment quality.

In contrast, the SAQ results generated by Wu-Palmer (WP) semantic similarity measure have the lowest SAQ values (rank of 15) compared to all other measures. In fact, all the other SAQ values are above 0.900. The WP SAQ measure indicates that that the alignment quality is not as good (around 0.7) across all OA systems and also has a larger variance of 0.004 compared to that of the other SAQ measures except for the JC measure which has a comparable variance of 0.0044.

To investigate why the WP measure produces the lowest SAQ values, the average semantic similarity value using WP are calculated between MA and HA concepts in all 1520*1519/2 reference alignment pairs. The WP averages for the MA and HA are 0.018 and 0.315, respectively. A possible explanation for such difference is the MA has a maximum depth of only 7 while the maximum depth of HA is 13. The MA also has a structure less complex than the HA's. Only 4% of its concepts have multiple parents; 13% of HA concepts have multiple parents. Smaller semantic similarity results for the MA for the WP measure could produce greater differences in the similarity between pairs of MA concepts and the similarity between corresponding HA pairs of concepts. This larger difference produces a larger $simDiff_{average}$ and therefore, the smaller SAQ values when the WP measure is used.

A clear ordering of the results exists for the semantic similarity measures in the SAQ. The downset SAQ measures produce the highest values. The path based measure WP produces the lowest SAQ values. The performance ordering of the OA systems for a specific semantic similarity measure, however, is not evident. For most of the measures, there is little difference in the SAQ values across the OA systems. Further analysis of the SAQ values in Table 2, Table 3a and Table 3b shows no semantic similarity measures except for the WP and JC have a difference between their maximum and minimum values of more than 0.01. The WP SAQ does have the lowest SAQ value for the GeRMeSMB OA system which corresponds with its ranking for the standard f-measure measure. From these results, however, the use of semantic similarity measures in this form of ontology alignment evaluation metric does not appear effective for assessing the quality of an ontology alignment.

VII. Conclusions and Future Plans

The research goal is to develop additional means of evaluating ontology alignment (OA) results that do not depend on a gold standard reference alignment. This research investigates the use of the lower cardinality distance measure in the discrepancy measure proposed in [26] for the OA evaluation task. It also presents a more general form of a semantic alignment quality (SAQ) measure. Experiments with

two real-world ontologies, the mouse anatomy and the NCIT human anatomy, used in the Ontology Alignment Evaluation Initiative (OAEI) demonstrate differences in the performance of a wide variety of semantic similarity measures in the general SAQ. These results led to further investigations that uncover some considerations on using a semantic distance or similarity measure in evaluating OA results.

Using the descendent set of a concept in assessing semantic distance or similarity between two ontology concepts is not a good selection for a feature set describing a concept. As indicated in this research, over 50% of the concept pairs for which distance or similarity was being measure, had no common descendents. The concepts, therefore, have no similarity even though they might indeed be siblings. Other results from the experiment indicate that as currently proposed the SAQ measure, regardless of the selected distance or semantic similarity measure, do not sufficiently distinguish among the performances of the OA systems as compared to the standard performance measures such as the f-measure.

Although the experimental results reveal problems with the SAQ measure, more research is needed to investigate ways of improving the SAQ measure for the evaluation task of OA results. Little research has been done to develop other methods of such evaluation that do not rely on a gold standard reference alignment. One idea that needs further study is that assessing the difference in the structure and composition of the two ontologies and factoring that into the SAQ evaluation might give evaluation results that correspond more with standard OA evaluation measures with a reference alignment.

As presented in section 3, only a few semantic similarity measures such as the Lin measure have been used in the OA process itself. Currently, work is underway to implement a general semantic similarity matcher in AgreementMaker in order to improve the OA process. Experiments may then be performed to determine how useful semantic similarity measures with mediating ontologies are to the OA task.

REFERENCES

[1] M. Ehrig, Ontology Alignment: Bridging the Semantic Gap, Springer Science+Business Media, LLC 2007.

[2] J. Euzenat, A. Ferrara, C. Meilicke, A. Nikolov, J. Pane, F. Scharfffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn, "The Results of the Ontology Alignment Evaluation Initiative 2010. Ontology Matching Workshop, International Semantic Web Conference, Shanghai, China 2010.

[3] J. Euzenat, "An API for ontology alignment," Proc of the 3rd International Semantic Web Conference (ISWC-2004), Hiroshima, Japan, 2004.

[4] N. Goodman, "Seven strictures on similarity" In N. Goodman (Ed.), Problems and projects. Pp. 437-447. Bobbs-Merrill, New York, 1972.

[5] D.L. Medin, R. L. Goldstone, and D. Gentner, "Respects for Similarity," *Psychological Review* 100(2), pp. 254-278, 1993.

[6] V. Cross and T. Sudkamp, Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications. New York: Physica-Verlag, ISBN 3-7908-1458, 2002.

[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, "Learning to Map between Ontologies on the Semantic Web," Proc. 11th Int 'l Conf. World Wide Web (WWW '02*),* pp. 662-673, 2002.

[8] M. Sabou, M. D'Aquin, E. Motta, "Exploring the Semantic Web as Background Knowledge for Ontology Matching," J. Data Semantics 11: pp. 156-190, 2008.

[9] M.A. Rodriguez and M. J. Egenhofer, "Determining Semantic Similarity among Entity Classes from Different Ontologies," IEEE Transactions on Knowledge and Data Engineering 15(2), pp 442-456, 2003.

[10] A.Tversky, "Features of Similarity," *Psychological Rev.*, vol. 84, pp. 327-352, 1977.

[11] A. Maedche and S. Staab, "Measuring Similarities Between ontologies," Proc. of the 13th European Conference on Knowledge Management and Acquisition (EKAW-2002, Siguenza, Spain, 2002.

[12] R. Rada, H. Mili E. Bicknell, M. Blettner, "Development and application of a metric on semantic nets," IEEE *Transaction on Systems, Man, and Cybernetics,* 19. pp 17–30, 1989.

[13] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in WordNet: An Electronic Lexical Database, *c.* Fellbaum, Ed. Cambridge, MA: MIT Press, pp. 265–283, 1998.

[14] Z. Wu and M. Palmer, "Verb semantics and lexical selection," Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp. 133-138, June 1994.

[15] P. Resnik, "Using information content to evaluate semantic similarity in taxonomy," Proc. Of the 14th Intl Joint Conference on Artificial Intelligence, pp, 448–453, 1995.

[16] N. Seco, N., T. Veale, J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet, In: ECAI. pp. 1089-1090, 2004.

[17] D. Lin, "An information-theoretic definition of similarity," Proc. of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann. pp 296–304, 1998.

[18] J. Euzenat and P. Valtchev, "An integrative proximity measure for ontology alignment," Proc.ISWC-2003 Workshop on Semantic Information integration, Sanibel Island (FL US), pp. 33–38, 2003.

[19] X. Su, Semantic enrichment for ontology mapping, Ph.D. Thesis, Computer & Information Science, Norwegian Univ of Science and Technology, 2004.

[20] J. Gracia, V. Lopez, M. d'Aquin, M. Sabou, E. Motta, and E. Mena, "Solving semantic ambiguity to improve semantic web based ontology matching. In: ISWC Workshop on Ontology Matching, 2007.

[21] V. Lopez, M. Sabou, and E. Motta, "Powermap: mapping the real semantic web on the fly," Proc. of 5th International Semantic Web Conference, Athens, GA, USA, November 5-9, 2006, LNCS 4273, 2006

[22] Y. R. Jean-Mary, E. P. Shironoshita, M. R. Kabuka, "Ontology matching with semantic verification," Web Semantics: Sci. Serv. Agents WorldWideWeb, 2009.

[23] G. Pirro, D. Talia, "UFOme: an ontology mapping system with strategy prediction capabilities," Data Knowl.Eng. 69.5, pp. 444-471, 2010,

[24] V. Cross, An analysis of fuzzy set aggregators and compatibility measures, Ph.D. Dissertation, Computer Science and Engineering, Wright State University, Dayton, OH, March 1993.

[25] V. Cross and X. Yu, "Investigating ontological similarity theoretically with fuzzy set theory, information content, and Tversky similarity and empirically with the gene ontology," Proc. of the 5th international conference on Scalable uncertainty management, 2011.

[26] C. A. Joslyn, P. Paulson, and A. White, "Measuring the Structural Preservation of Semantic Hierarchy Alignment," ISWC International Workshop on Ontology Matching. CEUR-WS, 2009.

[27] I.F. Cruz, C.Stroe, M. Caci, F Caimi, M. Palmonari, F P. Antonelli, and U.C. Keles, "Using AgreementMaker to Align Ontologies for OAEI." ISWC International Workshop on Ontology Matching (OM), ser. CEUR Workshop Proceedings vol. 689, pp. 118—125 2010.

[28] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," Proc. of the 10th International Conference on Research on Computational Linguistics, 1997.