



A Compliance Evaluation of a Hybridized Semantic Similarity Measure for Ontology Matching

Iroju Olaronke^{1*} and Ojerinde Oluwaseun²

¹Department of Computer Science, Adeyemi College of Education, Ondo, Ondo State, Nigeria.

²Department of Computer Science, Federal University of Technology, Minna, Niger State, Nigeria.

Authors' contributions

This work was carried out by both authors. Author IO designed the study, and wrote the first draft of the manuscript. Author OO managed the literature searches and participated in the statistical analysis of the study. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JSRR/2016/21593

Editor(s):

(1) Grigorios L. Kyriakopoulos, School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Greece.

Reviewers:

(1) Solehah Yaacob, International Islamic University, Kuala Lumpur, Malaysia.

(2) Dora Melo, Polytechnic Institute of Coimbra, Portugal.

(3) Sivakumar Ramakrishnan, Universiti Sains Malaysia, Malaysia.

Complete Peer review History: <http://sciencedomain.org/review-history/12057>

Original Research Article

Received 25th August 2015
Accepted 12th October 2015
Published 2nd November 2015

ABSTRACT

Semantic similarity measures are used to compute the common shared information between two concepts in domain resources such as ontologies. These measures form the basis of all ontology matching algorithms because they measure the degree of semantic relatedness between matched ontologies. There are two basic categories of semantic similarity measures used in ontology matching process. These measures include the structure based measure and the information based measure. The structure based measures deal with the structure of the ontology, that is, the shortest path length between two concept nodes in the ontology while the information based measures use the Information Content (IC) of concept nodes in the ontology. Despite the importance of semantic similarity measures in ontology matching, their accuracy still remains a challenge.

Aims: Consequently, this paper proposes a hybridized semantic similarity measure which combines the features of both structure and information based measures. The paper also evaluates the proposed measure against existing measures using compliance measures.

Methodology: The proposed measure was implemented using a subset of the Gene Ontology (GO) as the domain specific ontology. This was with the aim of measuring the semantic similarities

*Corresponding author: Email: irojuolaronke@gmail.com;

between muscular development terms and its hierarchies in the GO. The proposed measure was compared with Lin and Wu and Palmers' measures and evaluated on twenty one (21) pairs of the GO terms using precision, recall rate and F-measure.

Results: The result of the compliance evaluation showed that the proposed measure had a precision rate of 0.86, a recall rate of 1.00 and an F-measure of 0.92 as opposed to the Lin measure and the Wu and Palmer semantic similarity measures which had precisions of 0.83 and 0.79, recall rate of 0.53 and 0.61 as well as F-measures of 0.56 and 0.69 respectively.

Conclusion: The result of the compliance evaluation showed that the proposed measure performed better than Lin measure and Wu and Palmers' semantic similarity measure. However, the proposed measure needs to be evaluated using a larger data set in order to validate its reliability.

Keywords: Semantic similarity; ontology; ontology matching; compliance measures.

1. INTRODUCTION

The major goal of an ontology is to achieve a common and shared knowledge that can be transmitted between people and applications [1]. Consequently, ontologies play an imperative role in achieving interoperability by capturing domain knowledge and explicitly specifying semantics in a generic way [2]. However, diverse ontologies are developed for different applications, which result in semantic heterogeneity. Consequently, the need for matching ontologies to provide a common layer for information exchange arises.

Ontology matching processes typically involve analyzing and comparing ontologies to determine the semantic correspondences among their concepts by using semantic similarity measures. Semantic similarity measures, also referred to as semantic similarity techniques or approaches, estimate the semantic similarity between two hierarchically expressed concepts in a given ontology or taxonomy [3]. Semantic similarity measures can therefore be defined as a concept that is used to determine the degree to which two words, concepts or terms are semantically related based on an assigned metric. There are two basic semantic similarity techniques that are used to determine the semantic similarity between two or more ontologies. These measures include the structure based measures and the information based measures [3]. The structure based measures deal with the structure of the ontology, that is, the shortest path length between two concept nodes and/or depths of concept nodes in the ontology. This is based on the notion that the similarity of two concepts is higher when the two concepts have less distance between them. The information based measures on the other hand use Information Content (IC) of concept nodes derived from the IS-A relations in the ontology or taxonomy and corpus statistics.

However, the accuracy of these measures cannot be guaranteed [3].

This paper therefore appraises the structure based and information based semantic similarity measures used in ontology matching and their weaknesses. The paper also proposes a hybridized semantic similarity measure for ontology matching which exhibits the strengths of both the structure based measure and information based semantic similarity measure. The paper evaluates the performance of the proposed measure against existing measures based on compliance measures, specifically, precision, recall rate, F-measure.

The remainder of this paper is organized as follows: section 2 gives a brief overview of ontology, section 3 examines the concept of ontology matching while similarity measures in ontology matching is discussed in section 4. Compliance measures of ontology matching systems are described in section 5. The proposed semantic similarity measure is introduced in section 6 and implemented in section 7. Section 8 compares the proposed semantic similarity with existing semantic similarity measures using compliance measures. The limitation of the research is given in section 9, while the research is concluded in section 10.

2. AN OVERVIEW OF ONTOLOGY

Ontology is derived from the two Greek words namely ontos which means "to be" and logos which means "word" or "theory" [4]. According to Wu et al. [5], an ontology is a philosophical term which refers to a systematic approach that explains the existence of things in the world. An ontology provides a vocabulary that describes a domain of interest and a specification of the meaning of terms used in the vocabulary [6].

Ontology is composed of several components which include concepts, attributes, instances, axioms as well as relations. Concepts that are similar form Classes. A subclass is usually formed when classes are grouped into categories and when these categories have more specific categories below them. Hence, a Class C is a subclass of parent Class P if all of C's instances are also instances of P, which is the super class. Theoretically, this forms a taxonomy. A class may however have multiple super classes and subclasses. Attributes represent concrete data fields such as data type and properties, instances on the other hand are the concrete values of ontology concepts. Axioms restrict the possible interpretations of concepts, while relations represent the relationship that exists among two or more classes. Consequently, Zaib [7] described an ontology, O, as a 5-uple as shown in equations (1), (2), (3), (4), (5), (6), (7) and (8)

$$O = \langle C, A, I, R, D \rangle \tag{1}$$

Where C is the set of concept of the ontology,

$$C = \{c_1, \dots, c_k\} \tag{2}$$

c_1, \dots, c_k are the members of the concept C set while k is the number of members in concept C.

A is the set of attributes assigned to concept C

$$A = \{A(c_1), \dots, A(c_k)\} \tag{3}$$

$A(c_1), \dots, A(c_k)$ are the members of the set of attribute A

With

$$A(c_i) = \{a_{i1}, \dots, a_{in}\} \tag{4}$$

$A(c_i)$ is the set of attributes assigned to a concept c_i , while n is the number of attributes assigned to each of the concept.

R is the set of relations, r_1, \dots, r_m are members of set R and m is the number of members in the set R

$$R = \{r_1, \dots, r_m\} \tag{5}$$

With

$$r_p \in C \times C \times \sigma \tag{6}$$

r_p is the relation that connects two concepts with each other, σ denotes the natural alphabet in which the name/type of the relation is expressed.

I is the set of instance of the ontology, O, I_1, \dots, I_k are members of set I and k is the number of members in the set I

$$I = \{I_1, \dots, I_k\} \tag{7}$$

with

$$I_n = \{i_1, \dots, i_n\} \tag{8}$$

I_n is the set of instances assigned to a concept c_n , and

while D is a set of description logic sentences.

Ontology elements can be extracted from domain texts. This process is known as ontology learning. Ontology learning from texts can be subdivided into tasks based on the ontology element extracted from the resources. These tasks include term extraction, synonym extraction, concept extraction, relationship extraction, and axiom extraction [8].

2.1 Ontology Learning Approaches

There are different techniques that are employed during the process of ontology learning. These techniques include Natural Language Processing (NLP) techniques, dictionary based approaches, knowledge based approaches, semi structure data and relation schemas.

2.1.1 Natural language processing techniques

In general, Natural Language Processing techniques for ontology learning are categorized into three basic approaches, which include the symbolic, statistical, and hybrid approaches [9]. The symbolic approach utilizes linguistic components such as Noun Phrase chunking to extract information from text. For instance, noun phrases are considered to be linguistic representations of concepts and are often used to represent concepts in an ontology. The statistical approach utilizes large corpora of text data as well as statistical measurements to extract semantic information from texts. The hybrid approach however, is a combination of the symbolic and statistical approaches [9].

2.1.2 Dictionary based approach

This relies on machine readable dictionaries to extract relevant concepts and relations from texts [9].

2.1.3 Knowledge based approach

This relies on knowledge bases to extract relevant concepts and relations from texts [9].

2.1.4 Semi-structure data

This elicits data ontology from sources which have a predefined structure, such as Extensible Markup Languages (XML) schemas [9].

2.1.5 Relation schemas

This extracts relevant concepts and relations from knowledge in databases [9].

Ontologies are usually expressed in formal languages known as ontology languages. Examples of ontology languages include Web Ontology Language (OWL), Ontology Inference Layer (OIL), Description Logic, Topic Maps, Ontolingua and Resource Description Framework (RDF). Different ontologies can be developed for a particular domain using different tools and platforms such as Protégé, SWOOP and OntoEdit; they can as well be represented in different ontology languages. This introduces heterogeneity to ontologies.

2.2 Ontology Heterogeneity

There are diverse forms of ontology heterogeneity. According to Shaviko et al. [6], ontology heterogeneity includes terminological heterogeneity, semiotic or pragmatic heterogeneity, semantic heterogeneity and syntactic heterogeneity.

2.2.1 Terminological heterogeneity

This occurs when different terms represent the same concept in different ontologies. Example includes the terms heart and cardiac which represent synonymous terms.

2.2.2 Semiotic or pragmatic heterogeneity

This is usually caused as a result of subjective interpretation of terms by human beings. According to Zaib [7], terms with the same semantics can be interpreted in diverse ways. For instance, the abbreviation CP can represent diverse meanings such as credit points or commissioner of police.

2.2.3 Semantic heterogeneity

This occurs when different meanings are attached to a particular term.

2.2.4 Syntactic heterogeneity

This usually results when two or more ontologies are represented in different formal ontology languages such as RDF and OWL.

Hence, the need to match ontologies.

3. ONTOLOGY MATCHING

Ontology matching, according to Euzenat and Shvaiko [10], is a function which matches two input ontologies O_1 and O_2 by using a previous alignment A , a set of parameters, P , and thesauri, R . This process produces an alignment A' which represents the correspondences between the two input ontologies as described in equation (9).

$$A' = f(O_1, O_2, A, P, R) \quad (9)$$

The resulting alignment A' is defined as a set of correspondences which represent the relations between different entities. A correspondence according to Shvaiko [6] is a defined as:

Given two ontologies o and o' with associated entity languages Q_L and $Q_{L'}$, a set of alignment relations, θ , and a confidence structure over Ξ , a correspondence is a 5-uple:

$$\text{correspondence} = \langle id, e, e', r, n \rangle \quad (10)$$

such that

id is a unique identifier of the given correspondence;
 e and e' are entities of ontologies o and o' respectively

such that

$$e \in Q_L(o) \text{ and } e' \in Q_{L'}(o) ;$$

r is a relation

such that

$r \in \theta$
 n is a confidence measure which is usually in the range of $[0, 1]$
 such that
 $n \in \Xi$

The correspondence $\langle id, e, e', r, n \rangle$ asserts that the relation r holds between the ontology entities e and e' with confidence n .

In addition, ontology matching process is defined by Zaib [7] as shown in equations (11) and (12).

$$\text{Match: } (O, P) \rightarrow (c_1, c_2) \quad (11)$$

$$\text{sim}(c_1, c_2) > t \quad (12)$$

where

O and P are the ontologies,
 c_1 and c_2 are the concepts the ontology,
 such that
 $c_1 \in O$ and $c_2 \in P$
 $\text{sim}(c_1, c_2)$ is the similarity measure between
 the two entities
 t is the similarity threshold.

$$m(p_i) = \max\{A(n_c, n_{c'}) \times C(p_i, p_j)\}, \forall p_j \in P(c') \quad (14)$$

where

n_c and $n_{c'}$ denote the names of concepts c
 and c' ,
 such that $c \in C$ and $c' \in C$
 p_i and p_j denote the names of the properties
 of concept C
 such that $p_i \in P(c)$ and $p_j \in P(c')$
 A is the terminological affinity function.

Shallow matching, $SA_{c,c'}$ is evaluated as the
 weighted sum of the linguistic affinity of c and c'
 and their contextual affinity. This is described by
 Castano et al. [11] as shown in equation (15).

$$SA_{c,c'} = W_{la} \times A(n_c, n_{c'}) + \frac{(1-W_{la}) \sum_{i=1}^{|P(c)|} m(p_i)}{|P(c)|} \quad (15)$$

where

W_{la} is the weighted sum of the linguistic affinity
 n_c and $n_{c'}$ are the names of concepts c and c' ,
 $P(c)$ is the property of concept C
 A is the the terminological affinity function
 $m(p_i)$ is the result of the best matching

3.1 Types of Ontology Matching

Ontology matching according to Castano et al.
 [11] can be done in diverse ways. These include
 surface matching, shallow matching, deep
 matching and intensive matching.

3.1.1 Surface matching

This kind of ontology matching considers the
 linguistic features of concept descriptions.
 Surface matching deals with high- level and
 poorly structure ontological description. Given
 two concepts c and c' , surface matching provides
 a measure $SA_{c,c'}$, of their semantic affinity by
 exploiting the terminological affinity function [11].
 Surface matching, $SA_{c,c'}$ is defined by Castano et
 al. [11] as shown in equation (13)

$$SA_{c,c'} = A(n_c, n_{c'}) \quad (13)$$

where

n_c and $n_{c'}$ are the names of concepts c and c'
 respectively,
 $SA_{c,c'}$ is the semantic affinity between the
 concepts c and c' and
 A is the terminological affinity function.

3.1.2 Shallow matching

This kind of matching considers the concept
 names n_c and $n_{c'}$ and concept properties p_i and
 p_j . It also takes into consideration the linguistic
 features, the properties as well as the cardinality
 constraints of the ontologies. For the property
 comparison, each property $p_i \in P(c)$ is matched
 against all properties $p_j \in P(c')$. Furthermore, the
 best matching, $m(p_i)$, as described by Castano et
 al. [11] is as described in equation (14)

3.1.3 Deep matching

The deep matching model according to Castano
 et al. [11] considers concept names, and the
 whole context of concepts in terms of properties
 and semantic relations. The concept of deep
 matching is mathematically represented by
 Castano et al. [11] as shown in equation (16).

$$SA_{c,c'} = W_{la} \times A(n_c, n_{c'}) + (1 - W_{la}) \frac{\sum_{i=1}^{|Cxt(c)|} m(p_i)}{|Cxt(c)|} \quad (16)$$

where

W_{la} is the weighted sum of the linguistic affinity
 n_c and $n_{c'}$ are the names of concepts c and c' ,
 A is the the terminological affinity function
 $Cxt(c)$ is the whole context of the concepts in
 terms of properties and semantic relations
 $m(p_i)$ is the result of the best matching
 p_i denotes the name of the properties of concept
 C .

3.1.4 Intensive matching

The intensive matching model considers concept
 names, and whole context of concepts as well as
 the property values of the concepts [11]. The
 concept of intensive matching is mathematically
 represented by Castano et al. [11] as

$$SA_{c,c'} = W_{la} \times A(n_c, n_{c'}) + (1 - W_{la}) \frac{\sum_{i=1}^{|Cxt(c)|} m(p_i) + \sum_{j=1}^{|P(c)|} (p_j)}{|Cxt(c)| + |P(c)|} \quad (17)$$

where

W_{la} is the weighted sum of the linguistic affinity
 n_c and $n_{c'}$ are the names of concepts c and c' ,
 A is the the terminological affinity function
 $Cxt(c)$ is the whole context of the concepts in terms of properties and semantic relations
 $m(p_i)$ is the result of the best matching
 p_j denotes the name of the properties of the concept
 such that $p_j \in P(c)$

4. SIMILARITY MEASURES IN ONTOLOGY MATCHING

Similarity measures form the basis of all matching algorithms, because they measure the degree of similarity between two ontologies O_1 and O_2 to be matched. Zaib [7] describes the formal notation for similarity measure in ontology matching as follows:

$$sim: E \times E \rightarrow R \quad (18)$$

$$E = E_1 \cup E_2 \quad (19)$$

where

E_1 is the entity set of ontology O_1 ,
 E_2 the entity set of ontology O_2 .
 sim is the similarity value which must satisfy some properties.
 R is a value which is in the range of [0 1]

4.1 Properties of Similarity Measures

Zaib [7] emphasized that positiveness, maximality and symmetry are the properties of measures that are used to measure the similarity between two concepts in an ontology.

4.1.1 Positiveness

This implies that $\forall x \in E$ and $\forall y \in E$,

$$sim(x, y) \geq 0 \quad (20)$$

where

x and y might be concepts, labels or comments, numbers or sets of numbers or strings in the entity set E
 $sim(x, y)$ is the similarity value between x and y

4.1.2 Maximality

This implies that $\forall x \in E$ and $\forall y, z \in E$,

$$sim(x, x) \geq sim(y, z) \quad (21)$$

where

x, y and z might be concepts, labels or comments, numbers or sets of numbers or strings in the entity set E
 sim is the similarity value

4.1.3 Symmetry

This implies that $\forall x, y \in E$

$$sim(x, y) = sim(y, x) \quad (22)$$

where

x and y might be concepts, labels or comments, numbers or sets of numbers or strings in the entity set E
 sim is the similarity value

4.2 Typical Similarity Measures Similarity Measures in Ontology Matching

These include string based measures, token based distance, term frequency/inverse document frequency, language based methods, number measures and set based measures.

4.2.1 String based measures

The simplest possibility to determine the similarity of two strings is to test on equality. String equality is defined by Zaib [7] as:

$$string_{equ}(s, t) = \begin{cases} 0 & \text{if } s = t \\ 1 & \text{if } s \neq t \end{cases} \quad (23)$$

with

s and t being two strings
 $string_{equ}$ is the string similarity between s and t
 String equality returns "0" if the strings under consideration are identical and "1" if they are not identical.

4.2.2 Token-based distance

Mohammad [1] described the token-based distance as a technique that considers a string as a set of words or bag of words. The token-based distance considers a string as a vector s ,

which belongs to a metric space V , in which each dimension in the vector is a term or token, and each position in the vector is the number of occurrences of the token in the corresponding bag of words. This approach is used for comparing pieces of texts rather than labels. It usually works well on long texts.

4.2.3 Term frequency/inverse document frequency

This is widely used in the information community for determining the relevance of a term in a corpus by taking into account the frequency of the term in the corpus.

4.2.4 Language based methods

This is based on the use of natural language processing techniques to extract the meaningful terms from a text. Terms are usually phrases that identify concepts in ontologies. Shvaiko [6] emphasized that there are two general techniques of language based methods. These include the use of algorithms and the use of external linguistic resources such as dictionaries, thesauri, and terminologies. According to Shvaiko [6], a lexicon or dictionary comprises a set of words with their natural language definitions. A terminology is a kind of lexicon which contains phrases rather than single words. It is usually built for a specific domain, thus it is considered less equivalent than a dictionary. A thesaurus, on the other hand, is a kind of lexicon with more relational information.

4.2.5 Number measures

These are used to compare numbers in ontology matching process. Mathematically, number based measures according to Zaib [7] can be represented as follows:

$$\text{sim}_{\text{num}}(a, b) = \frac{1}{1+|(a-b)|} \quad (24)$$

where

a and b are two numbers
 sim_{num} is the similarity value between a and b

4.2.6 Set-based measures

This is useful when sets of values are considered for comparison. An example of the set based measure is the cosine similarity, which calculates the similarity of two vectors expressed as the

cosine of the angle between them. It is only applicable for vectors containing numbers. It is defined by Zaib [7] as:

$$\text{sim}_{\text{cos}}(v, w) = \frac{v \times w}{|v| \times |w|} \quad (25)$$

with

v and w being two vectors
 sim_{cos} is the cosine similarity between vectors v and w

For ontology matching purposes, the cosine similarity is always in the range of $[0, 1]$ because the values included in the vectors are always positive.

However, the above similarity measures do not measure the semantic relatedness or similarity between two or more ontology concepts [1].

Semantic Similarity Measures in Ontology Matching

Semantics represents meaning, while similarity is a complex concept which denotes relatedness, likeness or a common/shared feature or characteristics between two or more concepts. Thus, semantic similarity can simply be referred to as the degree to which two or more concepts are related in meaning. Islam and Inkpen [12] defined semantic similarity as a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning or semantic content. Danushka et al. [13] also viewed semantic similarity between two words as the posterior probability that they belong to the synonymous-words class. Semantic similarity according to Miller [14] also refers to the similarity between two concepts in taxonomy such as the Word Net, Cyc upper ontology and Unified Medical Language System Similarity. Semantic similarity can therefore be concisely defined as the degree of semantic likeness or relationship between two or more concepts in two or more ontologies based on their closeness in the ontology hierarchy. There are two basic ways of measuring the semantic similarity between two concepts in an ontology. These include the structure methods and the information based methods [15]. Structure based semantic similarity measures are usually based on the shortest path length and/or the depth of the concept nodes in the ontology [3]. Common structure based semantic similarity measures used in ontology matching include Wu and Palmer measure, Leacock and Chodorow and Li et al. semantic similarity measures, while the

information based measures are based on the Information Content (IC) of concept nodes in the ontology [3]. Examples of the information based measures include Resnik measure, Jiang and Conrath measure as well as the Lin measure [3]. These measures are explained below.

4.4 Structure Based Semantic Similarity Measures

Typical examples of the structure based semantic similarity measures include Wu and Palmer semantic similarity measure, Leacock and Chodorow measure and Li et al. semantic similarity measure.

4.4.1 Wu and Palmer semantic similarity measure

Wu and Palmer [16] semantic similarity measure determines how semantically similar two concepts are in a knowledge based resource such as an ontology, based on the depth of the two concepts in the resource and their lower common subsumer (LCS), that is the lowest node that can be a parent for the two concepts. Wu and Palmers algorithm is described by Sun et al. [17] as follows:

$$\text{sim}_{wp}(c_1, c_2) = \frac{2 \times \text{depth}(\text{Lso}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2) + 2 \times \text{depth}(\text{Lso}(c_1, c_2))} \quad (26)$$

Where

c_1 and c_2 are concepts or words
 $\text{depth}(c_1)$ is the shortest distance from the root to concept node c_1
 $\text{depth}(c_2)$ is the shortest distance from the root to concept node c_2
 $\text{depth}(\text{Lso}(c_1, c_2))$ is the shortest distance between concepts c_1, c_2
 The Wu and Palmers measure (Wup) is also written by Hoa [3] as:

$$\text{Wup} = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \quad (27)$$

where

c_1 and c_2 are concept nodes
 LCS node determines the common sharing of two concept nodes,
 depth is shortest path length between two concept nodes c_1 and c_2 .

However, the Wu and Palmer algorithm does not give more accurate values because it always considers the depth of the terms from the root node [18]. This implies that the Wu and Palmer algorithm is based on the assumption that within a conceptual domain, the semantic similarity of two concepts is defined by how closely they are related in the hierarchy of an ontology [19].

4.4.2 Leacock and chodorow similarity measure

Leacock and Chodorow similarity measure is based on the distance of two concepts c_1 and c_2 in a taxonomy/ontology [20]. Leacock and Chodorow similarity measure, l_{ch} , is given as:

$$l_{ch} = -\log(\text{length}(c_1, c_2) / (2 \times D)) \quad (28)$$

where

$\text{length}(c_1, c_2)$ is the length of the shortest path between the two concepts c_1 and c_2 and D is the maximum depth of the taxonomy.

The accuracy of this measure cannot also be guaranteed because it gives different semantic similarity result for two identical concepts [3].

4.4.3 Li et al. similarity measure

Li et al. semantic similarity measure considers the shortest path length, depth of the lowest common subsumer and the local density of two concept nodes in a taxonomy [21]. Li et al. semantic similarity measure, $s(w_1, w_2)$ is given as:

$$s(w_1, w_2) = e^{-a1} \times \frac{e^{b \times h} - e^{-b \times h}}{e^{b \times h} + e^{-b \times h}} \quad (29)$$

Where b is the shortest path lengths of the two concepts w_1 and w_2 ,

h is the depth between the two concepts.

Li et al. measure however has limitations that violate some of the intuitions and assumptions of ontology-based similarity as it gives different similarity values for different identical pairs of concepts [3].

4.5 Information Based Semantic Similarity Measures

The information-based semantic similarity measures are based on the information theory

which uses text corpus such as wordnet as secondary information source. They use the Information Content (IC) of concept nodes derived from the IS-A relations and corpus statistics to determine the similarity between two concepts [3]. Examples of the Information based semantic similarity measure include Resnik measure, Jiang and Conrath measure and Lin measure.

4.5.1 Resnik measure

Resnik [22] proposed an information-based statistic method for computing the semantic similarity between two or more ontology entities. The basic idea of this method is that the more information two concepts have in common, the more similar they are. Furthermore, Resnik [22] defines the notion of Information Content (IC) as a measure of the specificity of a given concept, and this is defined based on the probability of the occurrence of the concept in a large corpus or taxonomy. The concept of Resnik measure as described by Lin and Sandkuhl [23] is discussed as follows:

Let the taxonomy be augmented with the function $p:\sigma\rightarrow[0,1]$, such that for any concept $c \in \sigma$, $p(c)$ is the probability of encountering concept c in σ . If the taxonomy has a unique top node then its probability is 1. The Information Content of c according to Lin and Sandkuhl [23] is quantified as:

$$IC(c) = -\log p(c) \quad (30)$$

Where

$IC(c)$ is the Information Content of concept c
 $p(c)$ is the probability of the occurrence of the concept in a large corpus or taxonomy.
 then,

$$\text{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)] \quad (31)$$

where

c_1 and c_2 are concepts
 $S(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2 .
 The word similarity, sim , is also defined as follows:

$$\text{sim}(w_1, w_2) = \max_{c_1, c_2} [\text{sim}(c_1, c_2)] \quad (32)$$

where

$s(w_1)$ and $s(w_2)$ represent the set of concepts in the taxonomy that are senses of words w_1 and w_2 respectively c_1 overs $s(w_1)$, c_2 overs $s(w_2)$.

$\text{sim}(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2 .

The drawback of the Resnik measure is that it is dependent on the skeleton structure of the taxonomy and it also ignores the information of the taxonomy structure. Thus, the Resnik measure does not differentiate the similarity values of any pair of concepts in a sub-hierarchy as long as their lowest super-ordinate is the same [24].

4.5.2 Jiang and conrath measure

This is a modified form of Resnik measure. It measures the semantic similarity of two concepts by using the difference in the Information Content of the two concepts to indicate their similarity [25]. This measure returns a score denoting how similar two word senses are, based on the Information Content of the lowest common subsumer in the knowledge based resource and that of the input synsets. Mathematically, the Jiang and Conrath measure is described by Torres and Gelbukh [25] as follows:

$$\text{similarity}(C_1, C_2) = 2 \times IC(\text{LCS}(C_1, C_2)) - IC((C_1) + IC(C_2)) \quad (33)$$

where

C_1 and C_2 are concepts
 $\text{LCS}(C_1, C_2)$ is the lowest common subsumer between concepts C_1 and C_2
 IC is the Information Content of concept of the concepts C_1 and C_2
 $\text{similarity}(C_1, C_2)$ is the result of the Jiang and Conrath Measure

The Jiang and Conrath semantic similarity measure according to Torres and Gelbukh [25] has a low recall rate.

4.5.3 Lin measure

Lin measure is also a modified form of Resnik measure, which defines the similarity of two concepts as the ratio between the amount of information needed to state the commonality between the concepts and the information needed to fully describe them [26]. The mathematical representation of Lin measure as

described by Lin and Sandkuhl [23] is as shown in equation (34) below.

$$\text{sim}(c_1, c_2) = \frac{2 \times \text{IC}(c_3)}{\text{IC}(c_1) + \text{IC}(c_2)} \quad (34)$$

where

c_1 and c_2 are concepts

c_3 is the common subsumer with the highest IC of both c_1 and c_2

Lin measure uses Information Content of concept nodes and does not take into consideration the depth of the concept nodes and the lower super-ordinate.

5. COMPLIANCE MEASURES OF ONTOLOGY MATCHING SYSTEMS

Compliance measures evaluate the degree of compliance of a system with regards to a standard. They are used for computing the quality of the output provided by a system compared to a reference output. Examples of compliance measures according to Euzenat et al. [27] include precision, recall, fallout measure, F-measure as well as overall measure. Precision and recall are the most widely and commonly used measures in ontology matching systems. However, systems are often not comparable based on precision and recall only, because a system that has a higher recall has a lower precision and vice versa. Consequently, other compliance measures were introduced.

5.1 Precision

Precision is a common measure used in Information Retrieval to express the quality of a search result. Like in Information Retrieval, precision measures the ratio of correctly found correspondences, that is, true positives over the total number of returned correspondences. In logical terms, precision measures the correctness or accuracy of the matched ontologies. Precision can be enhanced by reducing the number of incorrectly matched terms. Given a reference alignment R and an alignment A, the precision $P(A, R)$ is defined as the number of correct found correspondences divided by the total number of found correspondences. Formally, precision is defined by Euzenat and Shvaiko [10] as:

$$P(A, R) = \frac{|R \cap A|}{|A|} \quad (35)$$

where

R is the reference alignment

A is the alignment

$P(A, R)$ is the precision

5.2 Recall

Recall is also a common measure in Information Retrieval [27]. It measures the ratio of correctly found correspondences, that is, true positives; over the total number of expected correspondences that is, true positives and true negatives. Given a reference alignment R, the recall of an alignment A is given by Euzenat et al. [27] as:

$$R(A, R) = \frac{|R \cap A|}{|R|} \quad (36)$$

where

R is the reference alignment

A is the alignment

$R(A, R)$ is the recall

5.3 Fallout Measure

The fallout measure measures the percentage of retrieved pairs which are false positives. Given a reference alignment R, the fallout of an alignment A is given by Euzenat et al. [27] as:

$$F(A, R) = \frac{|A| - |A \cap R|}{|A|} \quad (37)$$

where

R is the reference alignment

A is the alignment

$F(A, R)$ is the fallout measure

5.4 F-Measure

The F-measure is used to aggregate the result of precision and recall. Given a reference alignment R and a number α between 0 and 1, the F-measure of an alignment A is given by Euzenat et al. [27] as:

$$M(A, R) = \frac{P(A, R) \times R(A, R)}{(1 - \alpha) \times P(A, R) + \alpha \times R(A, R)} \quad (38)$$

where

$P(A, R)$ is the precision

$R(A, R)$ is the recall rate

$M(A, R)$ is the F-measure

α is any number between 0 and 1

If $\alpha = 1$, then the F-measure, $M(A,R)$, is equal to precision $P(A,R)$ and if $\alpha = 0$, the F-measure is equal to recall, $R(A,R)$. When the value $\alpha = 0.5$ is used, the harmonic mean of precision and recall is obtained. This is given by Euzenat et al. [27] as:

$$M_{0.5}(A, R) = \frac{2 \times P(A,R) \times R(A,R)}{P(A,R) + R(A,R)} \quad (39)$$

5.5 Overall Measure

The overall measure, according to Melnik et al. [28] is the ratio of the number of errors on the size of the expected alignment. Given a reference alignment R , the overall measure of an alignment A is given by Euzenat et al. [27] as:

$$O(A, R) = R(A, R) \times \left(2 - \frac{1}{P(A,R)}\right) \quad (40)$$

where

- $P(A,R)$ is the precision
- $R(A,R)$ is the recall rate
- $O(A, R)$ is the overallmeasure

6. THE PROPOSED SEMANTIC SIMILARITY MEASURE

The proposed semantic similarity measure determines the semantic relatedness between two concepts in an ontology based on the features of the structure based measure and information based measure. The proposed model considers the depth of the lower super-ordinate,

the shortest path length between the concepts as well as the Information Content of the concepts. Thus, the proposed measure complements the strengths of the existing measures which either takes into consideration the Information Content of the concept or the depths of the concepts. The proposed semantic similarity measure is defined as the ratio of the sum of shortest path between two ontology concepts $d(c_1,c_2)$ and the depth of the lower super-ordinate of the two concepts, $depth(Lso(c_1,c_2))$, to the sum of the Information Content of the two concepts. The proposed semantic similarity measure is mathematically represented in equation (41)

$$sim(c_1, c_2) = \frac{d(c_1,c_2) + depth(Lso(c_1,c_2))}{2 \times (IC(c_1) + IC(c_2))} \quad (41)$$

7. IMPLEMENTATION OF THE PROPOSED SEMANTIC SIMILARITY MEASURE

The proposed measure was implemented using a subset of the Gene Ontology (GO) as the domain specific ontology as shown in Fig.1., where T_1 to T_{13} represents the terms in the ontology. This was with the aim of measuring the semantic similarity between muscular development concepts and its hierarchies in the GO.

Table 1 shows the statistical information of the terms in the Gene Ontology sub graph as indicated by Naeem and Gillani [29].

Table 1. Statistical information of the terms in the subset of the gene ontology [29]

Term	GENE Term	Frequency (F)	Probability (P)	Information content (IC)
T_1	Digestive Tract Development	2	0.15	2.70
T_2	System Development	8	0.61	0.70
T_3	Tissue Development	2	0.15	2.70
T_4	Organ Development	3	0.23	2.11
T_5	Muscle structure development	2	0.15	2.70
T_6	Anatomical Structure Morphogenesis	2	0.15	2.70
T_7	Digestive System Development	2	0.15	2.70
T_8	Cardiovascular System Development	1	0.07	3.70
T_9	Circular System	2	0.15	2.70
T_{10}	Muscle Tissue Development	1	0.07	3.70
T_{11}	Muscle Organ Development	1	0.07	3.70
T_{12}	Organ Morphogenesis	1	0.07	3.70
T_{13}	Anatomical Structure Development	13	1	0.00

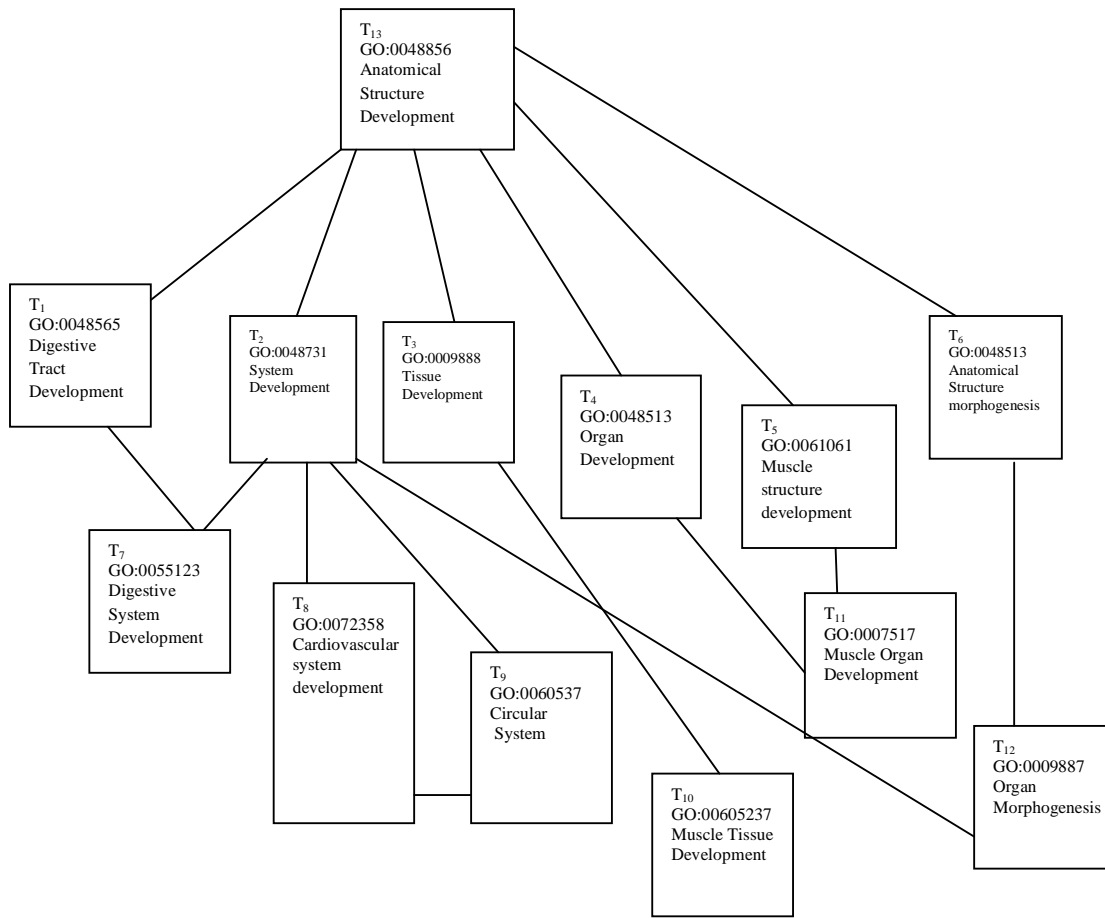


Fig. 1. Hierarchy of a subgraph of the gene ontology [29]

In Table 1 above, frequency, F , is defined as a measure of the number of descendants/children that a concept has including the concept itself. Probability, P , of a concept is defined as the ratio of the frequency of the concept to the maximum frequency as depicted in equation (41).

$$\text{Probability, } P(c) = \frac{F(c)}{\max(F(c))} \quad (42)$$

Where

$P(c)$ is the probability of a concept
 $F(c)$ is the frequency of the concept
 The IC is measured by taking logarithmic value of probability of a concept.

$$IC = -\log(P(c)) \quad (43)$$

where

$P(c)$ is the probability of concept c

$\log(P(c))$ is the logarithmic value of probability of the concept c

IC is the Information Content of concept c

Seven terms T_5 - T_{11} were purposively selected from the GO and matched using the proposed semantic similarity measure, Lin semantic similarity measure and Wu and Palmer semantic similarity measure. The terms were selected in a pair-wise manner. Since, there were seven terms to be matched, the statistical theory of combination shown in equation (44) was employed. Hence, a total of 21 pairs of the terms were matched.

$$C_2^m = \frac{m!}{2!(m-2)!} \quad (44)$$

where

m where m is the number of terms matched
 C_2^m is the statistical theory of combination

In Table 1, $\text{depth}(\text{Iso}(c_i, c_j))$ is the number of nodes that are between the two concepts c_i and c_j . For instance, there are two nodes between c_5 and c_7 in Fig. 1. These nodes are c_{13} and c_2 . In the same vein, there are three nodes between c_7 and c_{10} . These nodes are c_2 , c_{13} and c_3 . In addition, $\text{depth}(c_i)$ is the number of nodes between the root node c_i including c_i itself. For instance, the result of $\text{depth}(c_7)$ is 2. This was obtained by finding the number of nodes between c_{13} which is the root node and c_7 , which is one, that is c_2 , in addition to node c_7 itself. Similarly, $\text{depth}(c_j)$ is the number of nodes between the root node, c_{13} and c_j including c_j itself. The shortest path, $d(c_i, c_j)$ is the minimum number of edges between two concepts c_i and c_j . For instance, there are three edges between c_5 and c_7 . These are the edges connecting c_5 and c_{13} , c_{13} and c_1 as well as c_1 and c_7 in Fig. 1.

Table 3 shows the results of the semantic similarities between the pairs of terms matched using Lin measure, Wu and Palmer semantic similarity measure, and the proposed measure.

Recall that in equation (34) the Lin measure is given as:

$$\text{sim}(c_i, c_j) = \frac{2 \times \text{IC}(c_{\text{Isoh}})}{\text{IC}(c_i) + \text{IC}(c_j)} \quad (34)$$

where

c_i and c_j are concepts

c_{Isoh} is the lowest common subsumer, that is, the common node between c_1 and c_2 with the highest IC

For instance, for $\text{sim}(c_5, c_7)$ in Fig. 1,

$\text{IC}(c_5) = 2.70$

$\text{IC}(c_7) = 2.70$

Please refer to Table 1.

The lowest common subsumer between c_5 and c_7 is c_{13} and c_2 . However, c_2 has a higher IC of 0.70 than c_{13} with an IC of 0.00. Hence, the lowest common subsumer with the higher IC between c_5 and c_7 is c_2 . Therefore, $\text{IC}(c_{\text{Isoh}}) = 0.70$. Hence, $\text{sim}(c_5, c_7) = 1.00$

Similarly, recall that from equation (26), the Wu and Palmer's semantic similarity measure is given as:

$$\text{sim}_{\text{wp}}(c_1, c_2) = \frac{2 \times \text{depth}(\text{Lso}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2) + 2 \times \text{depth}(\text{Iso}(c_1, c_2))} \quad (26)$$

where

c_1 and c_2 are concepts

$\text{depth}(c_i)$ is the shortest distance from the root to concept node c_i ,

$\text{depth}(\text{Iso}(c_1, c_2))$ is the shortest distance between concepts c_1, c_2 .

For instance, to obtain the semantic similarity between concepts c_5 and c_7 $\text{sim}_{\text{wp}}(c_5, c_7)$ in Fig. 1.,

$\text{depth}(\text{Lso}(c_5, c_7)) = 2$,

$\text{depth}(c_5) = 1$

$\text{depth}(c_7) = 2$

Refer to Table 2. Hence, $\text{sim}_{\text{wp}}(c_5, c_7) = 0.57$

Also, recall that the proposed measure is given in equation (41) as:

$$\text{sim}(c_1, c_2) = \frac{d(c_1, c_2) + \text{depth}(\text{Lso}(c_1, c_2))}{2 \times (\text{IC}(c_1) + \text{IC}(c_2))} \quad (41)$$

where

$d(c_5, c_7) = 2$

$d(c_5, c_7) = 2$

Refer to Table 2.

Also,

$\text{IC}(c_5) = 2.70$

$\text{IC}(c_7) = 2.70$

Refer to Table 1. Hence $\text{sim}(c_5, c_7) = 0.28$

8. COMPLIANCE EVALUATION OF THE PROPOSED SEMANTIC SIMILARITY MEASURE

In order to evaluate the compliance of the proposed semantic similarity measure along with the existing measures, the knowledge of an expert in the Biomedical domain was sought in order to interpret the semantic similarity result in Table 3 and also to determine if the semantic similarity results returned by the three measures in Table 3 were accurate. Table 4 shows the human judgment on the proposed and existing semantic similarity measures.

In Table 4, semantic similarity scores below 0.6 are considered not semantically similar. However, semantic similarity score of 1.0 indicates that the terms are equivalent. As a rule of thumb, if the semantic similarity score is over 0.6, it means the terms are close matches [30]. Afterwards, the semantic similarity scores were evaluated using compliance measures specifically precision, recall rate as well as F-measure. The following formulas were employed in relation to Fig. 2.

Table 2. Pairs of matched terms, their shortest paths and the depth of their lowest super-ordinate

Term 1 (c _i)	Term 2 (c _j)	depth (c _i)	Depth (c _j)	depth (Lso (c _i , c _j))	d(c _i ,c _j)
T ₅	T ₆	1	1	1	2
T ₅	T ₇	1	2	2	2
T ₅	T ₈	1	2	2	3
T ₅	T ₉	1	2	2	3
T ₅	T ₁₀	1	2	2	3
T ₅	T ₁₁	1	2	0	1
T ₆	T ₇	1	2	2	3
T ₆	T ₈	1	2	2	3
T ₆	T ₉	1	2	2	3
T ₆	T ₁₀	1	2	2	3
T ₆	T ₁₁	1	2	2	3
T ₇	T ₈	2	2	1	3
T ₇	T ₉	2	2	1	3
T ₇	T ₁₀	2	2	3	4
T ₇	T ₁₁	2	2	3	4
T ₈	T ₉	2	2	1	2
T ₈	T ₁₀	2	2	3	4
T ₈	T ₁₁	2	2	3	4
T ₉	T ₁₀	2	2	3	4
T ₉	T ₁₁	2	2	3	4
T ₁₀	T ₁₁	2	2	3	4

Table 3. Result of the semantic similarities between the pairs of terms matched using Lin measure, wu and palmer semantic similarity measure, and the proposed measure

Term 1 (c _i)	Term 2 (c _j)	Lin	Wu and palmer	Proposed measure
T ₅	T ₆	0.00	0.5	0.28
T ₅	T ₇	1.00	0.57	0.37
T ₅	T ₈	0.22	0.57	0.39
T ₅	T ₉	0.26	0.57	0.46
T ₅	T ₁₀	0.22	0.57	0.39
T ₅	T ₁₁	0.65	0.00	0.08
T ₆	T ₇	0.26	0.57	0.46
T ₆	T ₈	0.22	0.57	0.39
T ₆	T ₉	0.26	0.57	0.46
T ₆	T ₁₀	0.84	0.57	0.39
T ₆	T ₁₁	0.65	0.57	0.39
T ₇	T ₈	0.22	0.29	0.31
T ₇	T ₉	0.26	0.29	0.37
T ₇	T ₁₀	0.84	0.67	0.55
T ₇	T ₁₁	0.65	0.67	0.55
T ₈	T ₉	0.22	0.29	0.23
T ₈	T ₁₀	0.73	0.67	0.47
T ₈	T ₁₁	0.57	0.67	0.47
T ₉	T ₁₀	0.84	0.67	0.55
T ₉	T ₁₁	0.65	0.67	0.55
T ₁₀	T ₁₁	0.75	0.67	0.55

$$\text{Precision, } P = \frac{TP}{TP+FP} \quad (44)$$

$$\text{Recall Rate} = \frac{TP}{TP+FN} \quad (45)$$

$$\text{F - measure, } F = \frac{2*(P*R)}{(P+R)} \quad (46)$$

where

TP is the True Positive,
 FP is the False Positive
 FN is the False Negative.

In Fig.2., the correct correspondences found by a matching system are called the True Positives (TP). This is as computed as shown in equation (47).

$$TP = S \cap H \quad (47)$$

Thus, a TP is obtained when a non semantically similar term, semantically similar term or equivalent term is correctly identified as non semantically similar term, semantically similar term or equivalent term respectively.

The incorrect correspondences found by a matching system are called the False Positives (FP). This is as shown in equation (48)

$$FP = S - S \cap H \quad (48)$$

A FP is therefore obtained when a non semantically similar term is incorrectly identified as semantically similar term or equivalent term.

The correct correspondences missed by a matching system are called False Negatives (FN). A FN is obtained when a semantically similar term or equivalent term is incorrectly identified as a non semantically similar term. This is as computed as shown in equation (48).

$$FN = H - S \cap H \quad (49)$$

where

$S \cap H$ represents the set of correspondences that are common to each of the semantic similarity measures and the human judgment.

Table 5 shows the TP, FP and FN for the Lin measure, Wu and Palmers' measure and the proposed measure.

Fig. 3 shows the graphical result of the compliance measure of proposed semantic similarity measure, Lin measure and Wu and Palmer's semantic similarity measure.

It can be deduced from Fig. 3 that the proposed semantic similarity measure has the highest precision, recall and F-measure values of 0.86, 1.00 and 0.92 respectively. In addition, the Lin measure gave a precision value of 0.83, a recall rate of 0.53 and F-measure of 0.56 while the Wu and Palmers Algorithm produced a precision value of 0.79, a recall rate of 0.61 and F-measure of 0.92.

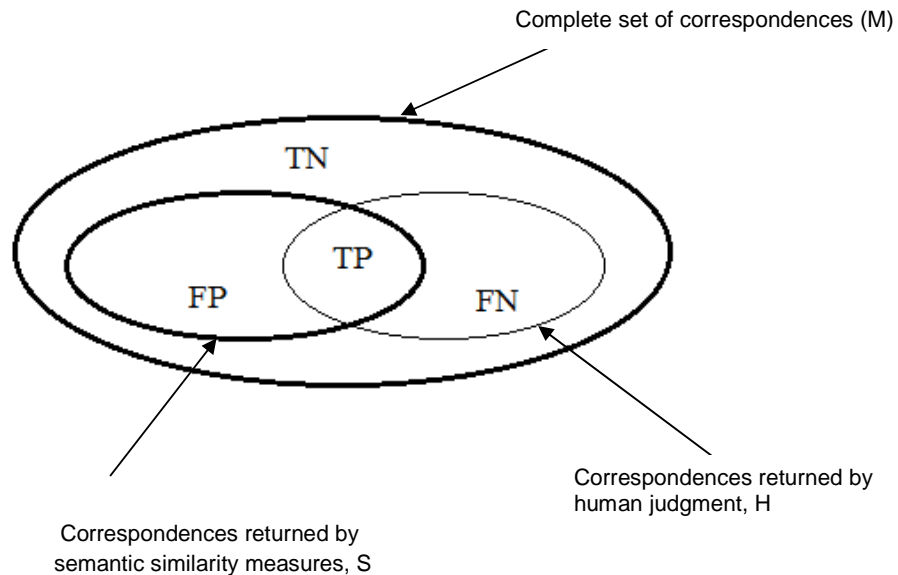


Fig. 2. Basic Set of Correspondences [32]

Table 4. A Comparative analysis of the proposed semantic similarity measure, existing measures and the expert judgment

Term 1 (c_i)	Term 2 (c_j)	Lin	Wu and palmer	Proposed	Human judgment
T ₅	T ₆	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₅	T ₇	equivalent terms	not semantically similar	not semantically similar	not semantically similar
T ₅	T ₈	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₅	T ₉	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₅	T ₁₀	not semantically similar	not semantically similar	not semantically similar	semantically similar
T ₅	T ₁₁	semantically similar	not semantically similar	not semantically similar	semantically similar
T ₆	T ₇	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₆	T ₈	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₆	T ₉	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₆	T ₁₀	semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₆	T ₁₁	semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₇	T ₈	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₇	T ₉	not semantically similar	not semantically similar	not semantically similar	not semantically similar
T ₇	T ₁₀	semantically similar	semantically similar	not semantically similar	not semantically similar
T ₇	T ₁₁	semantically similar	semantically similar	not semantically similar	not semantically similar
T ₈	T ₉	not semantically similar	not semantically similar	not semantically similar	equivalent terms
T ₈	T ₁₀	semantically similar	semantically similar	not semantically similar	not semantically similar
T ₈	T ₁₁	not semantically similar	semantically similar	not semantically similar	not semantically similar
T ₉	T ₁₀	semantically similar	semantically similar	not semantically similar	not semantically similar
T ₉	T ₁₁	semantically similar	semantically similar	not semantically similar	not semantically similar
T ₁₀	T ₁₁	semantically similar	semantically similar	not semantically similar	semantically similar

Table 5. TP, FN and FP of the semantic similarity measures

Correspondences	Lin	Wu and palmer	Proposed measure
TP	10	11	18
FP	2	3	3
FN	9	7	0

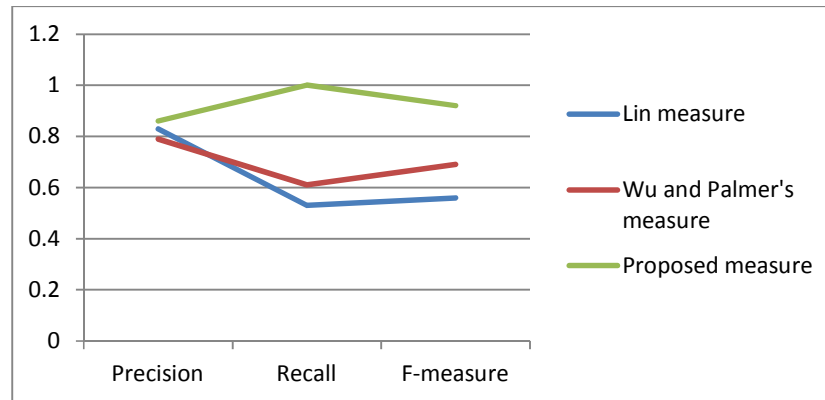


Fig. 3. Result of the compliance evaluation of lin measure, wu and palmer's measure and the proposed measure

9. LIMITATIONS OF THE STUDY

The proposed semantic similarity measure was tested on a small dataset of the Gene Ontology sub graph. However, in order to validate the reliability of the proposed semantic similarity measure, it needs to be evaluated on a larger data set. Furthermore, the proposed measure was compared with only two semantic similarity measures, each from the information based method (Lin Measure) and the structure based method (Wu and Palmers Algorithm). Nevertheless, there are diverse measures in both the structure based method such as Leacock and Chodorow similarity measure and the Li et al. measure and the information based measure such as the Jiang and Conrath semantic similarity measure and the Resnik measure.

10. CONCLUSION

Semantic similarity measures are very vital in the process of matching ontologies. This is because they measure the semantic relatedness between two or more ontologies. There are different similarity measures for measuring the semantic similarity between two or more concepts. These include the structured based measures such as the Wu and Palmer semantic similarity measure and the information based measures such as the Lin measure. The accuracy of these measures is however a challenge in the ontology matching process. Thus, this study proposes a hybridized semantic similarity measure that combines the features of both the structured based measures and the information based measures. This is a major distinctive feature between the proposed measure and existing measures. The proposed

measure takes into consideration the depth of the lower super-ordinate, the shortest path length between the concepts as well as the Information Content of the concepts. This is unlike existing measures which considers either the Information Content of the concepts or the depths of the concepts. The proposed measure was implemented on a sub graph of the Gene Ontology with the aim of measuring the semantic similarity between muscular development terms and its hierarchies in the ontology. Seven terms were purposively selected from the ontology; and these terms were matched in a pairwise manner. The statistical theory of combination was employed to determine the number of pairs of terms to be matched. A total of twenty one pairs of GO terms were matched using Lin measure, Wu and Palmer semantic similarity measure and the proposed measure. The result of the study showed that the proposed measure had a higher precision, recall and F-measure when compared with Lin measure and Wu and Palmer semantic similarity measure. The proposed measure recorded a precision of 0.86, a recall rate of 1.00 and an F-measure of 0.92 while the Lin measure gave of a precision of 0.83, recall of 0.53, and F-measure of 0.56. The Wu and Palmer semantic similarity measure had precision of 0.79, recall of 0.61 as well F-measures of 0.69. For future enhancement, this study will implement the proposed model on biomedical ontology tools such as Metamap and Concept Mapper on a larger data set.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Mohammad MT. State-of-the-art: ontology matching techniques and ontology mapping systems. *The International Journal of ACM Jordan*. 2010;1(3):84-68.
2. Mihoubi HS, Simonet, M. An ontology driven approach to ontology translation. *Proceedings of DEXA*. 2000;582-573.
3. Hoa AN. New semantic similarity techniques of concepts applied in the biomedical domain and wordnet. M.Sc. Thesis, University of Houston Clear Lake; 2006.
4. Iroju O, Soriyan A, Gambo I, Ikono R. State-of-the art: A comparative analysis of ontology matching systems. *African Journal of Computing & ICTs*. 2012;5(4): 91-79.
5. Wu YJ, Lang JS, Shang FH. A similarity-based approach for ontology mapping. *International Conference on Computer Science and Education*. 2009;169-165.
6. Shvaiko P. Iterative schema-based semantic matching. PhD Dissertation. International Doctorate School in Information and Communication Technology, University Of Trento; 2006.
7. Zaib SZ. Instance-based ontology matching and the evaluation of matching systems. Inaugural-Dissertation. Department of Computer Science, Heinrich Heine University of Dusseldorf, Germany; 2010.
8. Buitelaar P, Cimiano P, Magnini B. *Ontology learning from text: method, evaluation and applications*. IOS Press, Washington DC; 2005.
9. Lui K. *Ontology enrichment from free-text clinical documents: A comparison of alternative approaches*. PhD Thesis, School of Medicine. University of Pittsburgh; 2011.
10. Euzenat J, Shvaiko P. *Ontology matching*. Springer-Verlag, Heidelberg; 2007.
11. Castano S, Ferrara A, Montanelli S, Racca S. *Semantic information interoperability in open networked systems*. Università degli Studi di Milano, Italy; 2004.
12. Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transaction on Knowledge Discovery Data*. 2008;2(2):24-10.
13. Danushka B, Yutaka M, Mitsuru I. Measuring semantic similarity between words using web search engines. University of Tokyo, Hongo, Japan; 2008.
14. Miller GA. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*. 1990;3(4):264-245.
15. Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS'06*, New York, USA; 2006.
16. Wu Z, Palmer M. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*. 1994;138-133.
17. Sun KT, Huang YM, Liu MC. A wordnet-based near-synonyms and similar-looking word learning system. *Educational Technology and Society*. 2011;14(1):1341-21.
18. Vadivu G, Waheeta SH. Ontology mapping of Indian medicinal plants with standardized medical terms. *Journal of Computer Science*. 2012;8(9):1584-1576.
19. Al-Mubaid H, Nguyen HA. Using MEDLINE as standard corpus for measuring semantic similarity of concepts in the biomedical domain. *Proceedings of the IEEE 6th Symposium on Bioinformatics and Bioengineering BIBE-06*, Washington D.C., USA. 2006;319-315.
20. Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. *WordNet: An electronic lexical database*. MIT press. 1998;283-265.
21. Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*. 2003;15,4(2003):882-871.
22. Resnik P. Using information content to evaluate semantic similarity in ontology. *Proceedings of the 14th international Joint Conference on Artificial Intelligence*. 1995; 453-448.
23. Lin F, Sandkuhl K. A survey of exploiting word net in ontology matching. *International Federation for Information Processing, Artificial Intelligence and Practice II*. 2008;276:341.
24. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical ontology. *Proceedings of International*

- Conference on Research in Computational Linguistics. 1997;33-19.
25. Torres S, Gelbukh S. Comparing similarity measures for original wsd lesk algorithm. *Advances in Computer Science and Applications. Research in Computing Science.* 2009;43:166-155.
 26. Lin D. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning, San Francisco.* 1998;304-296.
 27. Euzenat J, Ehrig M, Raúl GC. Towards A methodology for evaluating alignment and matching algorithms version 1.0. *Ontology Alignment Evaluation Initiative;* 2005.
 28. Melnik S, Hector G, Erhard R. Similarity flooding: A versatile graph matching algorithm. *Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose, USA;* 2002.
 29. Naeem M, Gillani S. Gsemsim: Semantic similarity measure for intra gene ontology terms. *International Journal of Information Technology and Computer Science.* 2013; 06:40-32.
 30. Iroju OG, Soriyan HA, Gambo IP. Resolving semantic heterogeneity in healthcare: An ontology matching approach. *Journal of Computer Science and Engineering.* 2013;17(2):34-28.
 31. Giunchiglia F, McNeill F, Shvaiko P. Methodology for ontology matching quality evaluation. *Open Knowledge FP6-027253;* 2006.

© 2016 Olaronke and Oluwaseun; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:

*The peer review history for this paper can be accessed here:
<http://sciencedomain.org/review-history/12057>*