International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 341-356, 2018, ISSN:-2319–8656

Evaluating Semantic Similarity between Biomedical Concepts/Classes through Single Ontology

Abdelhakeem M. B. Abdelrahman
Sudan University of Science and Technology
Collage of Graduate Studies Khartoum, Sudan

Dr. Ahmad Kayed
Department of Computing and Information Technology
Sohar University, Sohar, Oman

Abstract Most of the existing semantic similarity measures that use ontology structure as their primary source can measure semantic similarity between concepts/classes using single ontology. The ontology-based semantic similarity techniques such as structure-based semantic similarity techniques (Path Length Measure, Wu and Palmer's Measure, and Leacock and Chodorow's measure), information content-based similarity techniques (Resnik's measure, Lin's measure), and biomedical domain ontology techniques (Al-Mubaid and Nguyen's measure (SimDist)) were evaluated relative to human experts' ratings, and compared on sets of concepts using the ICD-10 "V1.0" terminology within the UMLS. The experimental results validate the efficiency of the SemDist technique in single ontology, and demonstrate that SemDist semantic similarity techniques, compared with the existing techniques, gives the best overall results of correlation with experts' ratings.

Keywords: Biomedical information retrieval, biomedical ontology, semantic similarity measures, Unified Medical Language System (UMLS).

I. INTRODUCTION

Semantic similarity techniques are interested in measuring the semantic similarity, or inversely, semantic distance between two classes/concepts according to a given domain [8]. Semantic Similarity between two terms or sets of documents is defined as the degree of "sameness" between the terms as measured by comparing the information describing their properties [7]. Ontology-based semantic similarity measures are the similarity between two concepts, which is widely used in information retrieval and semantic web service fields [15]. They are can be roughly grouped into two groups as follows: 1) Ontology structure-based measures are those measures that use ontology taxonomy structure (is-a, part of) to calculate the similarity between concepts [5], [16]. In this measure the similarity between concepts is based on the path distance separating the concepts. These measures compute the similarity in terms of the shortest path between two concepts (classes) (group of synonyms) in the taxonomy. Rada et al, [5] proposed their measure as potential measure in the biomedical domain. Their experiments were conducted using MeSH (Medical Subject Headings) biomedical ontology. Wu and Palmer [16] proposed semantic similarity measure of concepts by taking into account the depth of concept nodes only. And 2) Information content-based similarity measures are those measures that use IC of concept derived from corpus statistics to measure the semantic similarity between concepts/classes. However, most of these semantic similarity measures can adapted to be use in biomedical domain. Hisham Al-Mubaid & Nguyen proposed new ontology-based semantic similarity measure that account for the depth of the concept nodes as well as distance (path length) between them. Another recent work on semantic similarity in biomedicine domain by Pedersen, Pakhomov and Patwardhan (2005) [7] in which they proposed a corpus-based context-vector approach to measure similarity between concepts in SNOMED-CT. Our contribution on this paper is compared between these semantic similarity techniques to choose the best measure among different similarity techniques that gives the best correlation and can be used to create our dataset or standard definition to be used to evaluate ontologies in the biomedical domain.

Most ontologies are developed for various purposes and domain [8]. For example, WordNet [8] is a lexical database for general English. In the biomedical domain, the Unified Medical Language System (UMLS) framework [8] includes many biomedical ontologies and terminologies (e.g., ICD-10, SNOMED-CT, MeSH, …etc).

II. BACKGROUND AND RELATED WORK

UMLS The Unified Medical Language System (UMLS) can be considered as an example of terminology which contains many clinical terms and integrates about 100 different vocabularies [1, 2]. It consists of three main knowledge sources: Meta thesaurus (MeSH, SNOMED-CT thesauruses, etc.), Semantic Network, and SPECIALIST Lexicon & Lexical Tools.

MeSH: MeSH, stands for Medical Subject Headings, [2, 3], is one of the source vocabularies used in UMLS. MeSH includes about 15 high-level categories, and each category is divided into subcategories and assigned a letter: A for Anatomy, B for Organisms and C for Diseases, and so on.

SNOMED-CT: SNOMED-CT, stands for Systemized Nomenclature of Medicine Clinical Term [2, 3], was included in UMLS in May 2004. It is a comprehensive clinical terminology, and the current version contains more than 360,000 concepts, 975,000 synonyms and 1,450,000 relationships organized into 18 hierarchies.

The following ontologies can be considered as known ontologies in the medical domain:

NCI Thesaurus (National Cancer Institute Thesaurus): an ontology vocabulary that includes broad coverage of the cancer domain, including cancer related disease, anatomy, genes and drugs.

ICD-10: [4], stand of International Classification Diseases version 10 is one of the most important international medical terminological systems; it was first issued in 1893. Its sixth revision was in 1948, and since this time it has been maintained by the World Health Organization (WHO). The current version is the tenth revision (ICD-10), which was issued in 1992. The initial aim of the ICD was to provide an international classification of death causes in order to produce internationally uniform and thus comparable mortality statistics. The WHO family of international classifications also includes other systems, notably the ICF (International Classification of Functioning, Disabilities and Health) and ICHI (International Classification of Health Inventions). The 22 main sub-categories of ICD-10 include, among others, diseases of the blood and blood-forming organs (D50–D89), endocrine, nutritional and metabolic diseases (E00–E90), mental and behavioral disorders (F00–F99), diseases of the nervous system (G00–G99) and certain infections and parasitic diseases (A00– B99). We present some preliminary observations about ICD-10 and consider the sub-domains I–XVII (codes A00 Q99). Core ontology of ICD-10 must explicate what sub-domains I–XVII address. Six of these domains are classified with respect to systems (nervous system, circulatory system, respiratory system, digestive system, musculo-skeletal system, genito-urinary system), three pertain to special organs (eye, ear, skin), and one domain relates to infectious diseases (A00–B99) and one domain addresses mental and behavioral disorders (F00–F99). Sub-domain level categories Level (i), i = I... XVII may be introduced; their instances are subsumed by the corresponding chapters. The instances of a level category level (i) in ICD-10 exhibit a taxonomic structure. Consider the domain of infections and parasitic diseases (A00–B99) and the associated domain-level category level (I), and includes about 21 high-level categories (taxonomies/sub trees) as shown in *Figure1*. The 2016 release of ICD-10 was used in our experiments.

For example as being described in figure1 below: In our experiment, the similarity is measured using different types of semantic similarity measures. From the evaluation result the best measure will be used in our benchmark dataset to evaluate ontologies in biomedical domain. Figure 1: below describes the biomedical domain type (ICD-10 ontology).

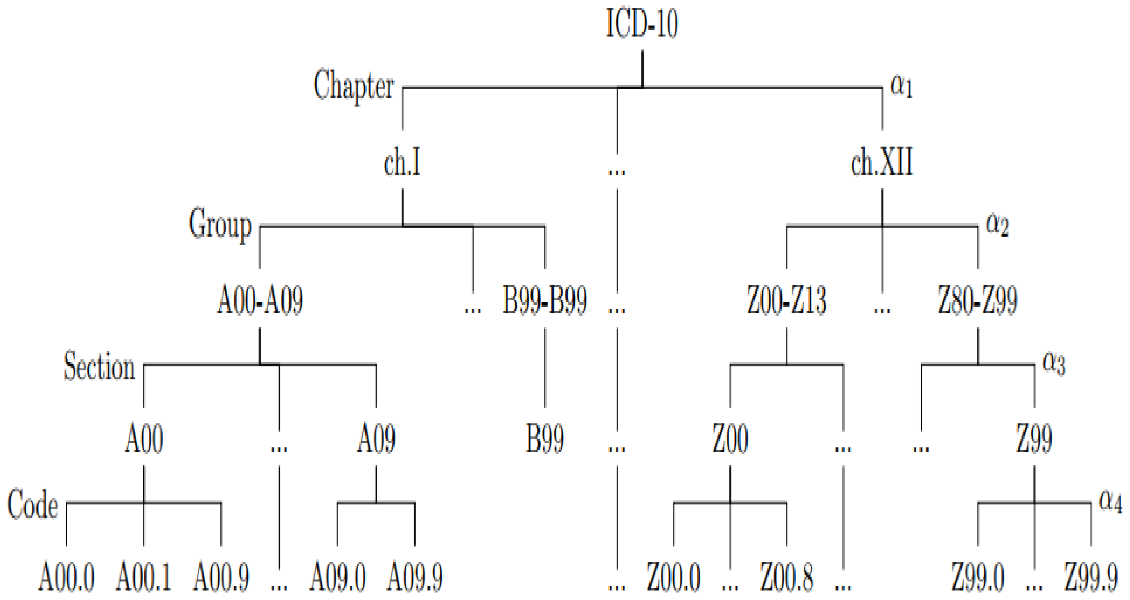


Figure 1: fragment of the ICD-10 taxonomy [17].

RELATED WORK

Rada et al. [5] [10] first proposed a semantic distance measure and applied it into the biomedical domain using MeSH ontology. The semantic distance between two classes is the shortest path length between them.

Caviedes and Cimino [6] [10] implemented the shortest Path length measure, called CDist, based on the shortest distance between two classes' nodes in the ontology. They evaluated their measure *(CDist measure)* on MeSH, SNOMED, ICD9 ontologies based on correlation with human ratings.

Pedersen et al. [7] [10] proposed semantic similarity and relatedness in the biomedicine domain in which they applied a corpus-based context vector approach to measure similarity between concepts in SNOMED-CT. Their context vector approach is ontology free but requires training text, for which, they used text data from Mayo Clinic corpus of medical notes.

Hisham Al-Mubaid & Nguyen [2] [8] proposed measure take the depth of their Least Common Subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concept are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure.

III. SEMANTIC SIMILARITY MEASURES

Semantic similarity techniques are becoming essential components of most of the information retrieval (IR), information extraction (IE), and other intelligent knowledge-based systems. For example, in IR, similarity measures play a crucial role in determining an optimal match between query terms and the retrieved document in ranking the results such as plagiarism detection [2]. The main semantic similarity measures could be classified into structure-based measures and information content (IC) measures.

1. Structure-based measures: In this measure the similarity between two concepts is based on the path distance separating the concepts. Which include the following types

1.1 Path Length Measure: finds the semantic distance between two concept nodes by finding the shortest path length between them on the ontology.

$$\text{Shortest Path}(C1, \quad C2) \quad = 2 * \text{Max}_{\text{depth}} - \text{len}(c1, c2) \qquad (1)$$

For example, to compute the similarity between "*Hypertensive renal disease with renal failure" (I12.0)* and "*Hypertensive renal disease with renal failure" (I12.0)* the shortest path length between them equal 1 "Using node counting"

Max_Depth of our Taxonomy = 5

So:

Sim (*Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure*) = 2*5 – 0 = 10 = 100%

.

Sim (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*) = 2*5 – 2 = 8= 20%

Table 1: Similarity values for two concepts from our taxonomy (Figure 1) using *Path Length* Based Measures (*shortest path*).

id	*Concept1*	*Concept2*	*LCA(c1 c2)*	*Length*	*Similarity*
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	0	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	*Heart failure I50*	2	80%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*ICD10_Chapter*	8	. . 20%

Rada et al. [64] estimates the distance of two concepts C_1 & C_2 as the shortest-path linking them $SP(C_1, C_2)$ and they used biomedical domain to evaluate their work in the information retrieval tasks using shortest path measure.

1.2 Wu and Palmer Measure: in this measure the similarity of concepts is compute by taking into account the depths of concept nodes only. They proposed a measure that has formula as follows:

$$\mathrm{Sim}\ (C1, C2) = 2 * \mathrm{depth}(LCS(C1, C2)) \ / \ (\mathrm{depth}(C1) + \mathrm{depth}(C2)) \quad (2)$$

The score can never be 0 because the depth of the LCS is never 0 (the depth of the root is 1) So the score is 0<Score<=1. When the two classes are the same the score is 1.

$$\mathrm{Sim}\ (C1, C2) = \frac{2N}{N1+N2+2N} \qquad (3)$$

Where N is the depth of the least common subsume (The least common subsume, LCS (C_1,C_2), of two concept nodes C_1 and C_2 is the lowest node that can be a parent for C_1 and C_2.

From our taxonomy (figure1), we can calculate the similarity between classes C1 and C2 as shown in table2:

Similarity (*Hypertensive renal disease with renal failure*, *Hypertensive renal disease with renal failure*) $= \frac{2*5}{0+0+(2*5)} = 1 = 100\%$

.

Similarity (*Pure hypercholesterolaemia*, *Lymph nodes of head, face and neck*) = $\frac{2*1}{4+4+(2*1)}$ = 0.2 = 20%

Table 2: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using Path Length Based Measures (*Wu & Palmer*).

id	Concept1	Concept2	LCS(c1 c2)	Wu & Palmer	Similarity
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	*Hypertensive renal disease with renal failure*	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	*Heart failure*	0.80	80%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*ICD10_Chapter*	0.20	20%

1.3 Leacock and Chodorow measure:

The similarity between two classes is determined by the shortest path length between two classes node, which connects these two classes in the taxonomy. The similarity is calculated as the negative algorithm of this value. They proposed a measure that has formula as follows:

$$\mathrm{SimL\&C} = -\log\left[\frac{\mathrm{Sp}\ (c1, c2)}{2(\mathrm{Max\_depth})}\right] \qquad (4)$$

Similarity (Hypertensive renal disease with renal failure, Hypertensive renal disease with renal failure) $= -\log\left(\frac{1}{2(5)}\right) = 1.00$

.

$$\text{Similarity (Congestive heart failure, Left ventricular failure)} = -\log\left(\frac{3}{2(5)}\right) \quad = 0.52287874528$$

.

$$\text{Similarity (Pure hypercholesterolaemia, Lymph nodes of head, face and neck)} = -\log\left(\frac{9}{2(5)}\right) \quad = 0.045757490560$$

Table 3: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using Path Leng1th Based Measures (Leacok and Chodorow).

ID	Concept1	Concept2	Length (c1 c2)	Leacok and Chodorow	Sim
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	3	0.52287874528	52%%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	9	0.0457574905607	. . 5%

2. Information Content-Based Similarity Measure:

2.1 Resnik's Measures

Resnik [9] the similarity between a pair of Classes (C_1 and C_2) is estimated as the amount of taxonomical information they share. In a taxonomy, this information is represented by the Least Common Subsume of both classes (LCS (C_1, C_2)), which is the most specific taxonomical ancestor common to C1 and C2 in a given ontology. Formally:

$$\text{Simres} = -\log(P(\text{LCS (C1, C2)}) = \text{IC}(\text{LCS (C1, C2)}) \qquad (5)$$

$$\text{IC(C)} = \frac{\log(\text{Depth(C)})}{\log(\text{Deep}_{\text{max}})} \qquad (6)$$

$$\text{IC}\big(\text{LCS(Hypertensive renal disease with renal failure,}$$
$$\text{Hypertensive renal disease with renal failure)}\big)$$
$$= \text{IC(Hypertensive renal disease with renal failure)}$$

Depth (Hypertensive renal disease with renal failure) = 5 "using node counting"

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC(Hypertensive renal disease with renal failure)} = \frac{\log(\text{depth(C)})}{\log(\text{deep}_{\text{max}})}$$

$$= \log\frac{(5)}{\log(5)} = 1.00$$

.

$$\text{IC}\big(\text{LCS(Congestive heart failure,} \quad \text{Left ventricular failure)}\big) = \text{IC(Heart failure)}$$

Depth (Heart failure) = 4 “using node counting”

Deepmax = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC}(\text{Heart failure}) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{deep}_{\text{max}})} = \log\frac{(4)}{\log(5)} = 0.86$$

.

$$\text{IC}\big(\text{LCS}(\text{Pure hypercholesterolaemia,} \quad \text{Lymph nodes of head, face and neck})\big)$$
$$= \text{IC}(\text{ICD10\_Chapter})$$

Depth (ICD10_Chapter) = 1 “using node counting”

Deep$_{\text{max}}$ = 5 the maximum depth of ICD10 Ontology.

Then:

$$\text{Simres} = \text{IC}(\text{ICD10\_Chapter}) = \frac{\log(\text{depth}(\text{C}))}{\log(\text{deep}_{\text{max}})} = \log\frac{(1)}{\log(5)} = 0.00$$

Table 4: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using information content based Measures (Resink).

ID	Concept1	Concept2	LCS(c1 c2)	SimResink	Similarity
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	5	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	4	0.86135311614	86%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	*1*	0.00	0.00%

2.2 Lin’s Measure:

This measure depends on the relation between information content (IC) of the LCS of two classes and the sum of the information content of the individual concepts [9].

$$\text{SimLin (c1, c2)} = \frac{2\times\text{IC(LCS (C1,C2))}}{\text{IC(C1)} + \text{IC(C2)}} \qquad (7)$$

From Resink’s Measure:

$$\text{IC}(\text{LCS}(\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure},$$
$$\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure}))$$
$$= \text{IC}(\mathit{Hypertensive\ renal\ disease\ with\ renal\ failure}) = \frac{\log(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Hypertensive\ renal\ disease\ with\ renal\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})}$$

$$= \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{SimLin}\ (Hypertensive\ renal\ disease\ with\ renal\ failure,$$
$$Hypertensive\ renal\ disease\ with\ renal\ failure) = \frac{2 \times 1}{1\ +\ 1} = 1.00$$

From Resink's Measure:

$$\mathrm{IC(LCS(}Congestive\ heart\ failure,\quad Left\ ventricular\ failure))$$
$$= \mathrm{IC}(Heart\ failure) = \frac{\log(4)}{\log(5)} = 0.86$$

$$\mathrm{IC}(Congestive\ heart\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Left\ ventricular\ failure) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{Sim}Lin(Congestive\ heart\ failure,\quad Left\ ventricular\ failure) = \frac{2 \times 0.86}{1\ +\ 1}$$
$$= 0.86$$

From Resink's Measure:

$$\mathrm{IC(LCS(}Pure\ hypercholesterolaemia\ ,\quad Lymph\ nodes\ of\ head, face\ and\ neck))$$
$$= \mathrm{IC(ICD10\_Chapter)} = \frac{\log(0)}{\log(5)} = 0.00$$

$$\mathrm{IC}(Lymph\ nodes\ of\ head, face\ and\ neck) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

$$\mathrm{IC}(Pure\ hypercholesterolaemia) = \frac{\log(\mathrm{depth(C)})}{\log(\mathrm{Deep_{max}})} = \log\frac{(5)}{\log(5)} = 1.00$$

Then:

$$\mathrm{Sim}Lin(ure\ hypercholesterolaemia,\quad Lymph\ nodes\ of\ head, face\ and\ neck)$$
$$= \frac{2 \times 0.00}{1\ +\ 1} = 0.00$$

Table 5: Similarity values for two concepts from the ICD-10 taxonomy (Figure 1) using information content based measures (*Lin*).

ID	Concept1	Concept2	IC(c1)	IC(c2)	IC(LCS(c1,c2))	Sim*Lin*
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1.00	1.00	1.00	100%
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	1.00	1.00	0.86	86%
. . 30	*Pure hypercholesterolaemia* (E78.0)	*Lymph nodes of head, face and neck (C77.0)*	1.00	1.00	0.0	0%

3. BIOMEDICAL DOMAIN SIMILARITY MEASURES

3.1 Hisham Al-Mubaid & Nguyen measure [2] [8] proposed measure take the depth of their Least Common Subsume (LCS) and the distance of the shortest path between them. The higher similarity arises when the two concept are in the lower level of the hierarchy. Classes that are more similar with have a lower similarity score than classes that are less similar with this measure. Their similarity measure is:

$$\text{SemDist}\ (C1,\ C2) = \log_2([L\ (C1,\ C2) - 1]^{\alpha} \times [\text{CSpec}(C1, C2)]^{\beta} + k) \qquad (8)$$

$$\text{CSpec}(C1, C2) = D - \text{depth}\left(\text{LCS}(C1, C2)\right) \qquad (9)$$

Where:

$\alpha > 0$ and $\beta > 0$ are contribution factors of two features (Path and CSpec).

Depth (LCS(C1, C2)) is depth of LCS(C1, C2) using node counting.

L(C1, C2) is shortest path length between the two concept nodes.

D is maximum depth of the taxonomy.

K is constant, and CSpec feature is calculated as in (9). We use logarithm function (inverse of exponentiation) for semantic distance (8), which is the inverse of semantic similarity.

To insure the distance is positive and the combination is non-linear, k must be greater or equal to one (k >= l). In this paper, k=l is used in experiments. When two concept nodes have path length of 1 (Path=l) using node counting (i.e., they are in the same node in the ontology), they have a semantic distance (SemDist) equals to zero (i.e. maximum similarity) regardless of common specificity feature.

The maximum value of this measure occurs when one concept is the left-most leaf node, and the other concept is the right-most leaf node in the tree. In ICD10 terminology the maximum value is $\log_2$ ([22-1]*[5-1] + 2) equal 6.4262647547. Therefore, the similarity distance values will be in [1.0000, 6.4262647547] in ICD10 terminology.

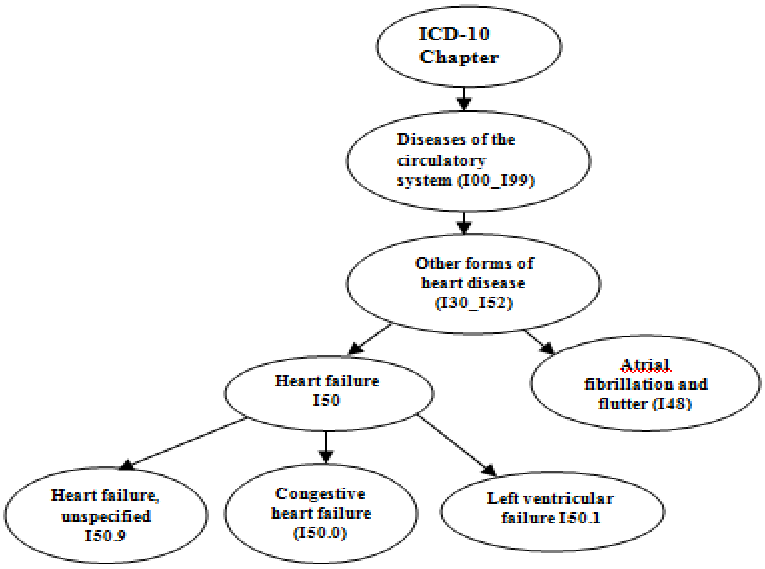

ICD-10 Chapter
Diseases of the circulatory system (I00_I99)
Other forms of heart disease (I30_I52)
Heart failure I50
Atrial fibrillation and flutter (I48)
Heart failure, unspecified I50.9
Congestive heart failure (I50.0)
Left ventricular failure I50.1

Figure 2. Hierarchy tree of eight concepts.

1) The single-cluster path length feature:

From our taxonomy (Figure 2), We can calculate the similarity between classes C1 and C2 as the following:

Path length (*Congestive heart failure, Left ventricular failure*) = 1 “using node counting”

CSpec (*Congestive heart failure, Left ventricular failure*) = D – depth (LCS (*Heart failure*))

= 5 – 4 = 1

So, similarity

Sim (*Congestive heart failure, Left ventricular failure*) = $\log_2([3 - 1]^1 \times [1]^1 + 2) = \log_2(4) = 2$

2) The cross-cluster path length feature:

Let us conceder the example, shown in Figure 3 below. The root is node that connects all the clusters. The path length between two concept nodes (C1 and C2) is computed by adding up the two shortest path lengths from the two nodes to their LCS node (their LCS is the root). For example, in Figure 1, for the two concept nodes *(Heart failure, unspecified, Atrial fibrillation and flutter*), the LCS is the root ICD-10. So, the path length between *Pure hypercholesterolaemia*, and *Lymph nodes of head, face and neck* is calculated as follows:

Path (*Pure hypercholesterolaemia, Lymph nodes of head, face and neck*) = d1 + d2 -1

Where d1 = d (*Pure hypercholesterolaemia*, root) and d2 = d (*Lymph nodes of head, face and neck*, root), where d (*Pure hypercholesterolaemia*, root) is the path length from the root ICD-10 to node *Pure hypercholesterolaemia*, and similarly d (*Lymph nodes of head, face and neck*, root) is the path length from ICD-10 to node *Lymph nodes of head, face and neck*. One is subtracted in the above equation, because the root node is counted twice.

$$\text{Path } (\mathit{Pure\ hypercholesterolaemia, Lymph\ nodes\ of\ head, face\ and\ neck}) = \text{d1} + \frac{2\text{D1} - 1}{2\text{D2} - 1} \times \text{ d2} - 1$$

$$\text{Path}(\mathit{Pure\ hypercholesterolaemia,\ \ Lymph\ nodes\ of\ head, face\ and\ neck})\ 5 + \frac{10-1}{10-1} \times\ 5 - 1\quad = 9$$

$$\text{CSpec (Pure hypercholesterolaemia, Lymph nodes of head, face and neck) = D primary} \ \ - 1 \ = \ 5 - 1 \ = 4$$

So, similarity

SemDist (Pure hypercholesterolaemia, Lymph nodes of head, face and neck)

$$= \log_2(\ [\text{Path} - 1]^{\alpha} \times [\text{CSpec}]^{\beta} + \text{k}) \ \ = \text{Log}_2\ ((9 - 1)\ \times\ (4) + 2) = \log_2 (34) = 5.09$$

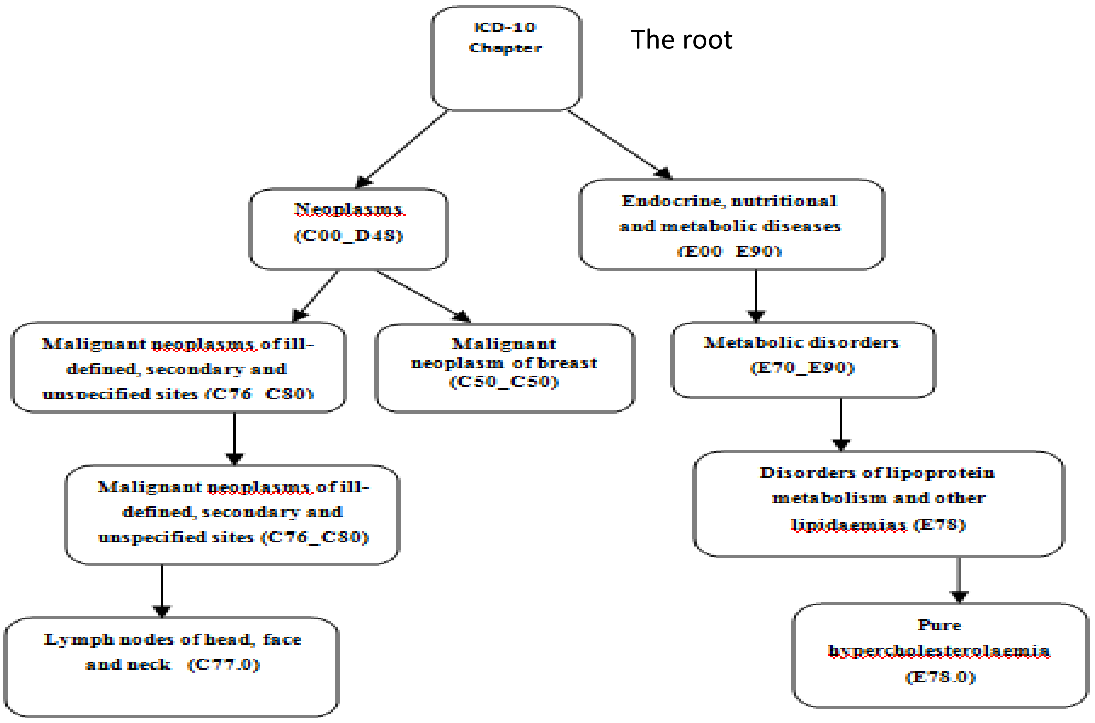

ICD-10 Chapter
The root
Neoplasms (C00_D48)
Endocrine, nutritional and metabolic diseases (E00 E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76 C80)
Malignant neoplasm of breast (C50_C50)
Metabolic disorders (E70_E90)
Malignant neoplasms of ill-defined, secondary and unspecified sites (C76_C80)
Disorders of lipoprotein metabolism and other lipidaemias (E78)
Lymph nodes of head, face and neck (C77.0)
Pure hypercholesterolaemia (E78.0)

Figure 3: Fragment of two clusters in ICD-10 Ontology (C77.0, E78.0).

Table 6: Similarity values for two classes from the ICD-10 taxonomy (Figure 1) using Path Length Based Measures (Al-Mubaid and Nguyen).

ID	Concept1	Concept2	L (c1,c2)	CSPec(c1, c2)	SimDist	Note
4	*Hypertensive renal disease with renal failure(I12.0)*	*Hypertensive renal disease with renal failure (I12.0)*	1	0	1	Same code
11	*Congestive heart failure* (I50.0)	*Left ventricular failure* (I50.1)	3	1	2	Same group
. . 30		*Lymph nodes of head, face and neck (C77.0)*	9	4	5.09	Different chapter

	Pure hypercholesterolaemia (E78.0)					

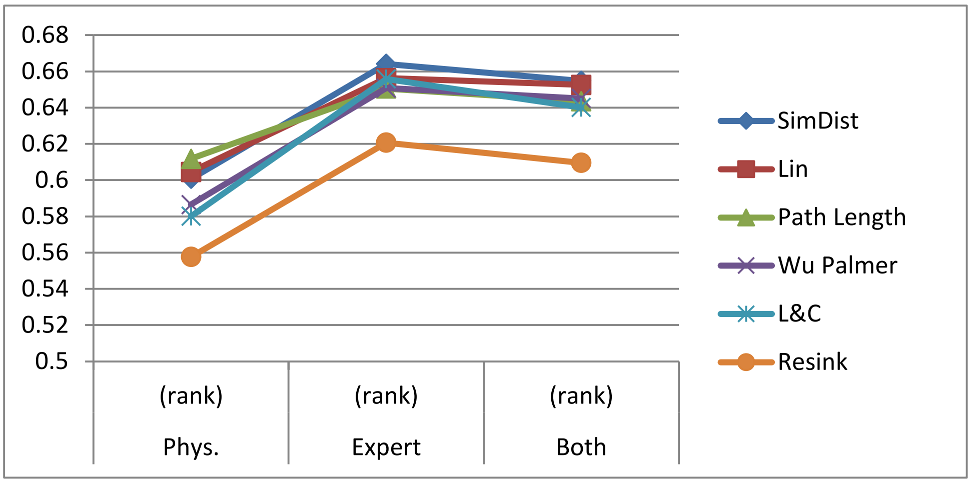
IV. EXPERMENTS AND RESULTS

For experiments, Ontologies of ICD10 were used as information source for the semantic similarity measure and one dataset are used for evaluation. All the measures use node counting for *path lengths* and *depths* of concept nodes. Out of the 30 pairs of Dataset 1 as shown in table 2, only 24 pairs in ICD10 were found. For the six pairs that were not found in ICD10 ontology, average distance/similarity values of the most related concept nodes to each one of them were calculated, so there were 24 pairs in ICD10 ontology in total. The results of absolute correlations with human scores using dataset1, experimented on ICD10 Ontology, are shown in Tables 7 and Figure 4. The experimental results demonstrated that *Al-Mubaid and Nguyen's measure (SimDist)* measure can achieve high correlations with human similarity scores.

Table 7: Absolute correlations with human scores for all measures using ICD10 on Dataset 1

Measure	**Phys.** **(rank)**	**Expert** **(rank)**	**Both** **(rank)**
SimDist	0.6007 (3)	**0.6641 (1)**	**0.6548 (1)**
Lin	0.6045 (2)	0.6563 (2)	0.6526 (2)
Path Length	**0.6118 (1)**	0.6505 (5)	0.6436 (4)
Wu Palmer	0.5865 (4)	0.6508 (4)	0.6451 (3)
L&C	0.5801 (5)	0.6558 (3)	0.6401 (5)
Resink	0.5576 (6)	0.6207 (6)	0.6096 (6)

Figure 4: Results of correlations with human scores for six measures using ICD10 Ontology.



V. EVALUATION:

Dataset:

There are no standard human rating sets of concepts/terms for semantic similarity in the biomedical domain. Thus, to evaluate the six semantic similarity measures, the dataset of 30 concept pairs from Pedersen et al. (2005) [7], (dataset1) which was annotated by 3 physicians and 9 medical index experts. Each pair was annotated on a 4-point scale: “practically synonymous, related, marginally related, and unrelated”.

Table 8 contains the whole pairs of this dataset. The average correlation between physicians is 0.68, and between experts is 0.78. Because the experts are more than the physicians, and the correlation (agreement) between experts (0.78) is higher than the correlation between physicians (0.68), it can be assumed that the experts’ rating scores are more reliable than the physicians’ rating scores.

Only 24 out of the 30 term pairs are found in ICD10 using ICD10 browser version 2010 [11] as some terms cannot be found, 24 pairs was used in the experiments (Pedersen et. al. [7] tested 29 out of the 30 concept pairs as one pair was not found in SNOMED-CT).

The term pairs in **bold**, in Table 8, are the ones that contains a term that was not found in ICD10 Ontology and they were excluded from experiments.

Table7 and figure4 show that the results of correlation with human ratings of physicians, experts, and both (phys. and experts), with the ranks between parentheses. These correlation values (Table7) show that the *SimDist* measure is ranked #1 in correlation relative to experts’ judgments and relative to both (expert and phys. judgments). But relative to physician judgments, the *SimDist* measure is ranked #3. From the applicability point of view, Nguyen and Al-Mubaid Measure (*SimDist*) is the most adequate one, and that can be used in our benchmark dataset. Finally the experiment describe above manually, should be obtained automatically. Hence, we need some software application or tools that can perform all the experiments automatically.

VI. CONCLUSION AND FUTURE WORK

The results discussed in this research has shown that, the SemDist (C1, C2) similarity (proposed by Al-Mubaid and Hoa A. Nguyen) has achieved high matching score by the expert's judgment to measure the similarity between two concepts in biomedical domain. In the future work of this research, we plan to implement a web-based system for all these semantic similarity measures and to make it available to researchers over the Internet.

Table 8 Dataset 1: 30 medical term pairs sorted in the order of the average.

Id	*Concept1*	*Concept2*	*Phys*	*Expert*	*Id*	*Concept1*	*Concept2*	*Phys*	*Expert*
4	Renal failure I12.0	Kidney failure I12.0	4.0000	4.0000	**27**	**Acne**	**Syringe**	**2.0000**	**1.0000**
5	Heart I51.5	Myocardium I51.5	3.3333	3.0000	12	Antibiotic (Z88.1)	Allergy (Z88.1)	1.6667	1.2222
1	Stroke I64	Infarct I64	3.0000	2.7778	**13**	**Cortisone**	**Total knee replacement**	**1.6667**	**1.0000**
7	Abortion O03	Miscarriage O03	3.0000	3.3333	**14**	**Pulmonary embolus**	**Myocardial infarction**	**1.6667**	**1.2222**
9	Delusion (F06.2)	Schizophrenia (F06.2)	3.0000	2.2222	16	Pulmonary Fibrosis (E84.0)	Lung Cancer (C34.1)	1.6667	1.4444
11	Congestive heart failure (I50.0)	Pulmonary edema (I50.1)	3.0000	1.4444	**6**	**Cholangiocarcinoma**	**Colonoscopy**	**1.3333**	**1.0000**
8	Metastasis (C77.0)	Adenocarcinoma (C08.9)	2.6667	1.7778	29	Lymphoid hyperplasia (K38.0)	Laryngeal Cancer (C32.0)	1.3333	1.0000
17	Calcification (M61)	Stenosis (H04.5)	2.6667	2.0000	21	Multiple Sclerosis (F06.8)	Psychosis (F06.8)	1.0000	1.0000
10	**Diarrhea**	**Stomach cramps**	**2.3333**	**1.3333**	22	Appendicitis (K35)	Osteoporosis (M80)	1.0000	1.0000
19	Mitral stenosis (I05.0)	Atrial fibrillation (I48)	2.3333	1.3333	23	Rectal polyp (K62.1)	Aorta (I70.0)	1.0000	1.0000
20	Chronic obstructive pulmonary disease (J44.9)	Lung infiltrates (J82)	2.0000	1.8889	24	Xerostomia (K11.7)	Alcoholic cirrhosis (K70.3)	1.0000	1.0000
2	Rheumatoid arthritis (M05.3)	Lupus (L93)	2.0000	1.1111	25	Peptic ulcer disease (K21.0)	Myopia (H52.1)	1.0000	1.0000
3	Brain tumor (G94.8)	Intracranial hemorrhage(I69.2)	2.0000	1.3333	26	Depression (F20.4)	Cellulitis (H60.1)	1.0000	1.0000
15	Carpal tunnel Syndrome (G56.0)	Osteoarthritis (M19.9)	2.0000	1.1111	**28**	**Varicose vein**	**Entire knee meniscus**	**1.0000**	**1.0000**
18	Diabetes mellitus (E10-E14)	Hypertension (I10-I15)	2.0000	1.0000	30	Hyperlipidemia (E78.0)	Metastasis (C77.0)	1.0000	1.0000

REFERENCES

[1] Hoa A. Nguyen, “New Semantic Similarity Techniques of Concepts Applied in the Biomedical Domain and WordNet” Master Thesis, Dec, 2006.

[2] Hisham Al-mubaid & Hoa A. Nguyen “Cluster-Based Approach for Semantic Similarity in the Biomedical Domain” Proceeding of the 28th IEEE, New York City, Aug 30-Sep 3, 2006.

[3] David Sánchez, “Semantic variance: An intuitive measure for ontology accuracy evaluation”, Engineering Applications of Artificial Intelligence 39 (2015) 89–99

[4] World Health Organization., “ICD-10, International Statistical Classification of Diseases and Related Health Problems.” 10th Revision. 5th ed. Vol.2 instruction manual (2016).

[5] Rada, et. al. “Development and Application of a Metric on Semantic Net”. IEEE Transactions on Systems, Man and Cybernetics, 19,1(1989),17-30.

[6] Caviedes, J. and Cimino, J. “Towards the development of a conceptual distance metric for the UMLS”. Journal of Biomedical Informatics 37,77-85, 2004.

[7] Pedersen,T. et al, “Measures of Semantic Similarity and Relatedness in the Medical Domain”, University of Minnesota Digital Technology Center Research Report, Journal of Biomedical Informatics April 2006.

[8] Hisham Al-mubaid & Hoa A. Nguyen “Measuring Semantic Similarity between Biomedical concepts within multiple ontologies” IEEE Trans Syst Man Cybern Part C: Appl Rev 2009, 39.

[9] Abdelrahman, A.M.B. and Kayed, A. (2015) A Survey on Semantic Similarity Measures between Concepts in Health Domain. American Journal of Computational Mathematics, 5, 204-214.

[10] Althobaiti, A.F.S. (2017) Comparison of Ontology-Based Semantic-Similarity Measures in the Biomedical Text. Journal of Computer and Communications, 5, 17-27.

[11] http://apps.who.int/classifications/icd10/browse/2010/en.

[12] Hliaoutakis, “Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline”. Master’s thesis, Technical University of Crete, Greek. 2005.

[13] UMLSKS. Available: http://umlsks.nlm.nih.gov

[14] MeSH Browser. Available: http://www.nlm.nih.gov/mesh/MBrowser.html

[15] Kaifeng, et. al. “AN IMPROVED METHOD FOR MEASURING CONCEPT SEMANTIC SIMILARITY COMBINING MULTIPLE METRICS”Proceedings of IEEE IC-BNMT2013.

[16] Wu, Z., and Palmer, M. Verb “semantics and lexical selection” 133-138, 1994.

[17] Roxana Dogaru, et. al, “Searching for Taxonomy-based Similarity Measures for Medical Data”BCI September 2015.