International Journal of Computer Applications Technology and Research
Volume 7–Issue 08, 331-340, 2018, ISSN:-2319–8656

Semantic Similarity Measures between Terms in the Biomedical Domain within frame work Unified Medical Language System (UMLS)

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

The techniques and tests are tools used to define how measure the goodness of ontology or its resources. The similarity between biomedical classes/concepts is an important task for the biomedical information extraction and knowledge discovery. However, most of the semantic similarity techniques can be adopted to be used in the biomedical domain (UMLS). Many experiments have been conducted to check the applicability of these measures. In this paper, we investigate to measure semantic similarity between two terms within single ontology or multiple ontologies in ICD-10 "V1.0" as primary source, and compare my results to human experts score by correlation coefficient.

Keywords: Information extraction, biomedical domain, semantic similarity techniques, Unified Medical Language System (UMLS), and Semantic Information Retrieval (SIR).

1. INTRODUCTION

Ontology is test bed of semantic web, capturing knowledge about certain area via providing relevant concept and relation between them. Quality metrics are essential to evaluate the quality. Metrics are based on structure and semantic level. At the present the ontology evaluation is based only on structural metrics, which has not been very appropriate in providing desired results.

Semantic similarity measures are widely used in Natural Language Processing. We show how six existing domain-independent measures can be adapted to the biomedical domain. Semantic similarity techniques are becoming important components in most intelligent knowledge-based and Semantic Information Retrieval (SIR) systems [1]. Measures and tests are provided to define how we can measure the "goodness" of ontology or its resources. Many experiments have been conducted to check the applicability of these measures [4].

General English ontology based structure similarity measures can be adopted to be used into the biomedical domain within UMLS. New approach for measuring semantic similarity between biomedical concepts using multiple ontologies is proposed by Al-Mubaid and Nguyen [2, 3]. They proposed new ontology structure based technique for measuring semantic similarity between single ontology and multiple ontologies in the biomedical domain within the frame work of Unified Medical Subject Language System (UMLS). Their proposed measure based on three features [2]: first Cross modified path length between two concepts. Second, new features of common specificity of concepts in the ontology. Third Local ontology granularity of ontology cluster.

2. BIOMEDICAL DOMAIN ONTOLOGIES

Most of the semantic similarity techniques work in the biomedical domain uses only ontology (e.g. MeSH, SOMED-CT) for computing the similarity between the biomedical terms[9]. However, in this work we use ICD- 10 ontology as primary source to computing the similarity between concepts in biomedical domain.

International Classification of Diseases (ICD): The newest edition (ICD- 10) is divided into 22 chapters: (Infections, Neoplasm, Blood Diseases, Endocrine Diseases, etc.), and denote about 14,000 classes of diseases and related problems. The first character of the ICD code is a letter, and each letter is associated with a particular chapter, except for the letter D, which is used in both Chapter II, Neoplasm, and Chapter III, Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism, and the letter H, which is used in both Chapter VII, Diseases of the eye and adnexa and Chapter VIII, Diseases of the ear and mastoid process. Four chapters (Chapters I, II, XIX and XX) use more than one letter in the first position of their codes. Each chapter contains sufficient three-character categories to cover its content; not all available codes are used, allowing space for future revision and expansion. Chapters I–XVII relate to

diseases and other morbid conditions, and Chapter XIX to injuries, poisoning and certain other consequences of external causes. The remaining chapters complete the range of subject matter nowadays included in diagnostic data. Chapter XVIII covers Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. Chapter XX, External causes of morbidity and mortality, was traditionally used to classify causes of injury and poisoning, but, since the Ninth Revision, has also provided for any recorded external cause of diseases and other morbid conditions. Finally, Chapter XXI, Factors influencing health status and contact with health services, is intended for the classification of data explaining the reason for contact with health-care services of a person not currently sick, or the circumstances in which the patient is receiving care at that particular time or otherwise having some bearing on that person's care [8, 10].

3. SEMANTIC SIMILARITY TECHNIQUES CHALLENGES IN THE BIOMEDICAL DOMAIN

Most of existing semantic similarity techniques that used ontology structure as the primary source can't measure the similarity between terms using single ontology or multiple ontologies in the biomedical domain within frame work Unified Medical Language System (UMLS). However, some of the semantic similarity techniques have been adopted to biomedical domain by incorporating domain information extracted from clinical data or medical ontologies.

4. RELATED WORK

4.1 Rada et al. Proposed semantic distance as a potential measure for semantic similarity between two concepts in MeSH, and implemented the shortest path length measure, called CDist, based on the shortest distance between two concept nodes in the ontology. They evaluated CDist on UMLS Metathesaurus (MeSH, SNOMED, ICD9), and then compared the CDist similarity scores to human expert scores by correlation coefficients.

4.2 Caviedes and cimino. [11] Implemented shortest path based measure, called CDist, based on the shortest distance between two concepts nodes in the ontology. They evaluated CDist on UMLS Metathesaurus (MeSH, SNOMED, ICD9), and then compared the CDist similarity scores to human expert scores by correlation coefficient.

4.3 Pedersen et al.[1] Proposed semantic similarity and relatedness in the biomedicine domain, by 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.

4.4 Wu and Palmer Similarity Measure [11] proposed a new method which define the semantic similarity techniques between concepts C_1 and C_2 as

$$\text{Sim}(C_1, C_2) = 2 \times \frac{N_3}{N_1 + N_2 + 2 \times N_3} \quad (1)$$

Where

N_1 is the length given as the number of nodes in the path from C_1 to C_3 which is the least common super concept of C1 and C2, and

N2 is the length given in the number of nodes on a path from C2 to C3.

N3 represents the global depth of the hierarchy and it serves as the scaling factor.

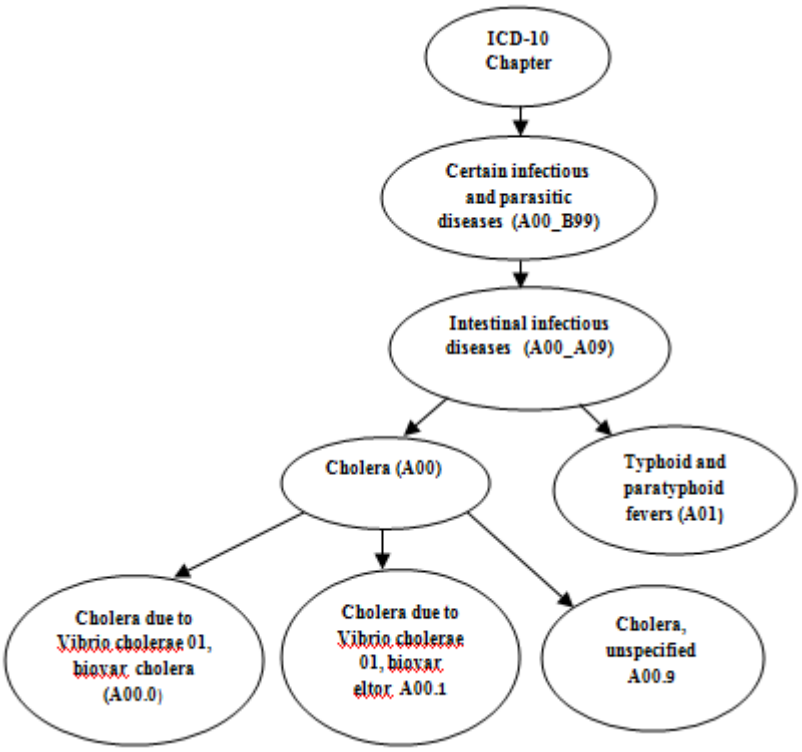

Figure 1 fragment of Intestinal infectious diseases

For example from Figure 1: (LCS (A00.1, A00.9) = A00 and LCS(A00 ,A01) = A00_A09) of two concept nodes and N_1, N_2 are the path lengths from each concept node to LCS, respectively. 4.5 Al- Mubaid and Nguyen Similarity technique [5, 11] 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 concepts are in the lower level of the hierarchy. Their similarity measure is:

$$Sim(c_1, c_2) = \log_2([L(c_1, c_2) - 1] \times [D - depth(L(c_1, c_2)] + 2) \quad (2)$$

Where:

$L(c_1, c_2)$ is the shortest distance between c1 and c_2.

Depth $L(c_1, c_2)$ is depth of $L(c_1, c_2)$ using node counting.

$L(c_1, c_2)$ lowest common subsume of c_1 and c_2.

D is the maximum depth of the taxonomy.

The similarity equal 1, where two concepts nodes are in the same cluster/ontology. The maximum value of this measure occur when one of the concepts is the left most leaf node, and the other concept is the right leaf node in the tree. In the ICD-10 tree let us consider an example in ICD-10 terminology. The category tree is “Intestinal infectious diseases” and is assigned letter A in ICD10 terminology version 2016 at the link (http://apps.who.int/classifications/icd10/browse/2016/en#/A00-A09). This tree looks as follows:

Intestinal infectious diseases [A00-A09]

Cholera [A00]+

Typhoid and paratyphoid fevers [A01]+

Other salmonella infections [A02]+

Shigellosis [A03]+

Viral and other specified intestinal infections [A08]+

Other gastroenteritis and colitis of infectious and unspecified origin [A09]+

The similarity between “Cholera [A00]” and “Typhoid and paratyphoid fevers [A01]” is less similarity than the similarity between “Cholera due to Vibrio cholerae 01, biovar eltor [A00.1]” and “Cholera, unspecified [A00.9]”. However, in this measure they take into account the depth

The symbol "+" indicates that the concept can be further expanded into a sub tree (sub-concepts). For example, "Cholera" [A00] can be expanded to be as follows:

Cholera [A00]

Cholera due to Vibrio cholerae 01, biovar cholerae [A00.0]+

Cholera due to Vibrio cholerae 01, biovar eltor [A00.1]+

Cholera, unspecified [A00.9]+

of the LCS of two concepts, in the path length and leacock & chodorwo produce semantic similarity for two pairs [(A00, A01) and (A00.1, A00.9)] in sim (c_1, c_2) measure (Eq 2 in table 1) give high similarity in lower level in the ontology hierarchy ([A00.1, A00.3]).

Table 1: Measures Comparison

Pair of Concepts	P. L	L. C	C. K	Hisham Al-Mubaid & Nyguan Measure (Eq 2)
A00 – A01	0.37	2.13	0.91	3.2
A00.1 – A00.9	0.33	2.15	0.91	1.6

The higher numeric similarity result between (A00, A01) means the lower semantic similarity between them.

5. EVALUATION

5.1 Datasets:

There are no standard human rating sets for semantic similarity in biomedical domain. Thus, Hisham Al-Mubaid and Nguyen [3, 11] used dataset from Pedersen et. al [1], which was annotated by 3 physician and 9 medical index experts to evaluate their proposed measure in the biomedical domain.

Table 2 Dataset 1: 30 medical term pairs sorted in the order of the average [1].

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

5.2 Experiments and Results

Table 2. Test set of 30 medical term pairs sorted in the order of the averaged physicians' scores (taken from Pedersen et. al. 2005 [1]). Al-Mubaid and Nguyen [5, 11] find only 24 out of the 30 concept pairs in ICD-10 using http://apps.who.int/classifications/icd10/browse/2016/en browser version 2010.

Another biomedical dataset was used containing 36 MeSH term pairs [15]. The human scores in this dataset are the average evaluated scores of reliable doctors. UMLSKS browser was used [12]

for SNOMED-CTterms, and MeSH Browser [13] for MeSH terms. Table 3, Table 4, Table 5, and Table 6 show Dataset2 along with human scores and scores of Path length, Wu and Palmer's, Leacock and Chodorow's, and Hisham Al-Mubaid & Nguyen techniques calculated using MeSH ontology. The term pairs in bold, in Table 3, Table 4, Table 5, and Table 6, are the ones that contain a term that was not found in MeSH Ontology and they were excluded from experiments.

Table3. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Path length's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Path length
1	Anemia	Appendicitis	0.031	8
2	Meningitis	Tricuspid Atresia	0.031	8
.		.	.	.
.		.	.	.
36	Chicken Pox	Varicella	0.968	1

Table 4. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Wu and Palmer's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Wu &Palmer
1	Anemia	Appendicitis	0.031	0.364
2	Meningitis	Tricuspid Atresia	0.031	0.364
.		.	.	.
.		.	.	.
36	Chicken Pox	Varicella	0.968	1.000

Table 5. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Leacock and Chodorow's scores using MeSH ontology.

Id	Concept 1	Concept 2	Human	Leacock & Chodorow
1	Anemia	Appendicitis	0.031	1.099
2	Meningitis	Tricuspid Atresia	0.031	1.099
.		.	.	.
36	Chicken Pox	Varicella	0.968	3.178

Table 6. Biomedical Dataset 2 (36 pairs) with human similarity scores (Human) and Hisham Al-Mubaid & Nguyen measure (SemDist) using MeSH ontology.

Id	Concept 1	Concept 2	Human	SemDist
1	Anemia	Appendicitis	0.031	4.263
2	Meningitis	Tricuspid Atresia	0.031	4.263
36	Chicken Pox	Varicella	0.968	0.000

6. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the basics of semantic similarity techniques, the classification of single ontology similarity measures and cross ontologies similarity measures. We prepare a brief introduction of the various semantic similarity measures in biomedical domain. However, from all the above, we can used SemDist as semantic similarity measures in the biomedical domain. In future work, we intend to explore the semantic similarity techniques in the biomedical domain (ICD10, MeSH, and SNOMED-CT) within UMLS frame work. We also prepare implement a web-based user interface for all these semantic similarity techniques and to make it available freely to researchers over the Internet. That will be much helpful for interested researchers in the field of bioinformatics text mining.

7. REFERENCES

[1] Ted Pedersen, et al. " Measures of semantic similarity and relatedness in the biomedical domain ", Journal of Biomedical Informatics 40 (2007) 288–299.

[2] Hisham Al-Mubaid and Hoa A. Nguyen, “A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain” Proceedings of the 28th IEEE, EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.

[3] 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.

[4] Ahmad Kayed, et al. "Ontology Evaluation: Which Test to Use" 2013 5th International Conference on Computer Science and Information Technology (CSIT), IEEE, pp 45-48, 2013.

[5] Hisham Al-Mubaid and Hoa A. Nguyen, "New Ontology Based Semantic Similarity for the Biomedical Domain", (2006) p 623 – 628.

[6] S. Anitha Elavarasi, et. al, "A Survey on Semantic Similarity Measure" International Journal of Research in Advent Technology, Vol.2, No.3, March 2014 E-ISSN: 2321-9637.

[7] Nguyen H., Al-Mubaid H. (2006) "New Semantic Similarity Techniques of Concepts applied in the biomedical domain and WordNet." MS Thesis, University o f Houston Clear Lake, Houston, TX USA, 2006.

[8] World Health Organization, "International statistical classification of diseases and related health problems". - 10th revision, edition 2010.

[9] Hisham Al-Mubaid and Hoa A. Nguyen, "Using MEDLINE as Standard Corpus for Measuring Semantic Similarity in the Biomedical Domain", Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06), 2006.

[10] Mirjana Ivanovic& Zoran Budimac, An overview of ontologies and data resources in medical domains, Expert Systems with Applications 41 (2014) 5158–5166.

[11] Montserrat Batet Sanromà, "ontology-based semantic clustering", PhD Thesis, 2010.

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

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