Concept Name Similarity Measure on SNOMED CT

Htet Htet Htun and Virach Sornlertlamvanich

School of Information, Computer and Communication Technology,
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
Email:htethtethtun.8910@gmail.com,virach@siit.tu.ac.th

Abstract. The semantic similarity measure between biomedical terms or concepts is a crucial task in biomedical information extraction and knowledge discovery. Most of the existing similarity approaches measure the similarity degree based on the path length between concept nodes as well as the depth of the ontology tree or hierarchy. These measures do not work well in case of the "primitive concepts" which are partially defined and have only few relations in the ontology structure. Namely, they cannot give the desired similarity results against human expert judge on the similarity among primitive concepts. In this paper, the existing two ontology-based measures are introduced and analyzed in order to determine their limitations with respect to the considered knowledge base. After that, a new similarity measure based on concept name analysis is proposed to solve the weakness of the existing similarity measures for primitive concepts. Using SNOMED CT as the input ontology, the accuracy of our proposal is evaluated and compared against other approaches with the human expert results based on different types of ontology concepts. Based on the correlation between the results of the evaluated measures and the human expert ratings, this paper analyzes the strength and weakness of each similarity measure for all ontology concepts.

Keywords: Concept Name Similarity Measure, Text Similarity, Natural Language Processing, SNOMED CT, Semantic Similarity

1 Introduction

Over the years, the determination of semantic similarity between word pairs has been recognized as an important task of text understanding applications such as the proper exploitation, management or classification of textual data [1], information retrieval [2] and decision-support systems that utilize knowledge sources and ontologies [3]. Semantic Similarity measures exploit knowledge sources as the base to perform the estimation. In recent years, knowledge sources and ontologies are widely used for the semantic similarity research area as they offer a structured and unambiguous representation of knowledge in the form of conceptualizations interconnected by means of semantic pointers. At the same

time, finding the semantic similarity between concepts based on the medical ontologies becomes crucial for the biomedical domain. As an example, health decision support system retrieves similar treatment cases in the past based on their different similarity levels as guidelines in order to treat the current patient [4]. Therefore, many ontology-based similarity measures have been developed by exploiting the medical ontologies such as SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms). SNOMED CT is considered as a standard medical terminology [5] that covers all areas of clinical information including body structure, diseases, organisms and clinical findings etc.

The fundamental idea of computing the similarity between words/concepts is based on the taxonomic structure of an ontology by taking the minimum number of shortest path between evaluated two concepts. Leacock and Chodorow [6] proposed a measure that considers both the shortest path length between two concepts and the maximum depth of the taxonomy but performs a logarithmic scaling. Wu and Palmer [8, 7] also proposed a path-based measure that takes into account the depth of the two concepts in the hierarchy and also the depth from their least common subsumer (LCS) to the root of the ontology. Choi and Kim [9] also proposed an taxonomic-based measure according to the difference in the depth levels of two concepts and the distance of the shortest path between them. AI-Mubaid and Nguyen [10] proposed a cluster-based measure that combines path length and common specificity by subtracting the depth of their LCS from the depth of the ontology. These previous measures use an ontology as background knowledge and determine the similarity based on taxonomic structure of an ontology by taking shortest path length and depth of evaluated concepts. But the new ontology taxonomic-based measure [11] analyzed previous measures and presented as taking only the minimum path length between evaluated concepts omits a large amount of taxonomic knowledge of the ontology for the evaluated concepts and waste a lot of explicitly available knowledge. Therefore, they proposed a new taxonomic-based measure by taking all possible number of parent concepts and they got the highest correlation results among previous measures against human expert ratings. In the literature, SNOMED CT is chosen as the input ontology to find the semantic similarity. When we analyze the SNOMED CT ontology, there are two kinds of concepts "defined concepts" and "primitive concepts" according to the available hierarchical information of the concepts in the ontology [12]. For example, definition of primitive concept "Tumor of dermis" is as follows.

`(Tumor of dermis` $\sqsubseteq$ `special concept)` $\sqcap$ `(special concept` $\sqsubseteq$ `SNOMED CT concept)`

It means "Tumor of dermis" has "is-a" relation with "special concept" and "special concept" has "is-a" relation with the top "SNOMED CT concept". The definition of another primitive concept "Vibrio species n-z' is the same with "tumor of dermis".

`(Vibrio species n-z` $\sqsubseteq$ `special concept)` $\sqcap$ `(special concept` $\sqsubseteq$ `SNOMED CT concept)`

Therefore, their definitions are the same and not sufficient to distinguish from each other and they needed to be additionally defined with the specific information in the ontology. For the "defined concepts", they are sufficiently defined in the ontology as follows.

$(\texttt{Hypoxia of brain} \equiv \texttt{Hypoxia} \sqcap \exists \texttt{FindingSite. Brain Structure})$

"Hypoxia of brain" has "is-a" relation with "Hypoxia" and it has "attribute-value" relation type "findingSite" with another concept "brain structure". Therefore, "defined concepts" have specific and complete information in the ontology but "primitive concepts" have partially hierarchical information and they are actually needed to define with complete information [12]. Therefore, released ontology versions include different amounts of concepts and relations in every year because ontology builders always redefine the concepts with more complete and specific information from the actual medical records. For this reason, ontology builders call the concepts that have been completely defined in the ontology as "defined concepts" and other concepts that are needed to add specific hierarchical information and relations in the ontology as "primitive concepts". In the literature, there is no experiments based on the different types of concepts for the SNOMED CT ontology. These facts push us to measure the similarity between different types of concepts. Therefore, we consider three cases of experiment for measuring the degree of similarity (1) between the primitive concepts, (2) between primitive concepts and defined concepts, and (3) between the defined concepts.

In this paper, we first review taxonomic-based measure [11] that got the highest correlation with the human expert result among most of the previous ontology-based measures in the literature. From the logic point of view, SNOMED CT definitions are written by Description Logic EL family, therefore, we review the description logic ELH similarity measure [13]. Then, we make the detail analysis between three cases of experiment by identifying limitations of previous measures according to dimensions of expected accuracy. In order to overcome the limitations of previous measures, we propose a new similarity measure based on concept name for all types of concepts. To compare all measures in a practical setting, we evaluate them against human expert results. The results show that our proposal solves the limitation of existing measures.

The rest of the paper is organized as follows. Section 2 presents and analyzes previous ontology-based similarity measures. Section 3 presents our similarity measure and its main benefits. Section 4 evaluates all measures based on three cases of experiments using SNOMED CT as the domain ontology and makes the comparison of evaluated measures with human expert ratings. Section 5 presents the conclusions.

2 Ontology-based Similarity Measures

2.1 Taxonomic-based (Path-based) Measure

This measure computes the similarity based on the taxonomic paths connecting the two concepts [11]. It considers all of the superconcepts belonging to all the

possible taxonomic paths between concepts. This relation is based on the idea that pairs of concepts belonging to an upper level of the taxonomy (i.e. they share few superconcepts) should be less similar than those in a lower level (i.e. they have more superconcepts in common). It defines the similarity between concept c_1 and c_2 as the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge, and then it takes the inverted logarithm function as shown in Eq. (1).

$$sim(c_1, c_2) = -log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \tag{1}$$

In the full concept hierarchy H^c of concepts (C) of an ontology, T(c_i) = { $c_j \in$ C — c_j is superconcept of c_i } ∪ { c_i } is defined as the union of the ancestors of the concept c_i and c_i itself.

This measure takes into account all the superconcepts regarding the evaluated concepts and it relies on the taxonomic paths. According to the SNOMED CT ontology, all of the ontology concepts are not well defined and completely defined especially for all primitive concepts. So similarity measure based on taxonomic structure of an ontology will be the main problem for the primitive concepts because they are actually needed to define with full hierarchical information from all actual medical records. As a consequence, similarity degree between primitive and defined concepts may not get the correct similarity value because primitive concepts have few inter-links and defined concepts have full inter-links between them so their similarity may be low but their actual similarity according to the judgment of human expert will be high.

2.2 ELSIM Similarity Measure

The ELSIM semantic similarity measure computes the similarity between $\mathcal{ELH}$ concepts based on homomorphism tree function. This function provides a numerical value that represents structural similarity of one concept description against the another concept description. This measure is used for description logic $\mathcal{ELH}$ definitions. It first constructs description tree for each concept from Top to evaluated concept using Algorithm 1 (see in detail [13]). Secondly, it maps between two description trees using homomorphism degree function as the following.

Definition 3.1 (Homomorphism Degree) Let $T^{\mathcal{ELH}}$ be a set of all $\mathcal{ELH}$ description trees and $\mathcal{T}_C$, $\mathcal{T}_D \in T^{\mathcal{ELH}}$ corresponds to two $\mathcal{ELH}$ concept names C and D, respectively [13]. The homomorphism degree function hd: $T^{\mathcal{ELH}} \times T^{\mathcal{ELH}} \rightarrow$ [0,1] is inductively defined as follows:

$$hd(\mathcal{T}_C, \mathcal{T}_D) = \mu \,.\, p - hd(\mathcal{P}_C, \mathcal{P}_D) + (1 - \mu) \cdot \text{e-set-hd}(\mathcal{E}_C, \mathcal{E}_D), \tag{2}$$

where $0 \leq \mu \leq 1$;

$$p - hd(\mathcal{P}_C, \mathcal{P}_D) := \begin{cases} 1 & if \mathcal{P}_C = \emptyset \\ \frac{|\mathcal{P}_C \cap \mathcal{P}_D|}{|\mathcal{P}_C|} & otherwise, \end{cases} \tag{3}$$

$\mathsf{e\text{-}set\text{-}hd}(\mathcal{E}_C, \mathcal{E}_D) :=$

$$\begin{cases} 1 & if \mathcal{E}_C = \emptyset \\ 0 & if \mathcal{E}_C \neq \emptyset \ and \ \mathcal{E}_D = \emptyset \\ \sum_{\epsilon_i \in \epsilon_C} \frac{max\{e-hd(\epsilon_i,\epsilon_j):\epsilon_j \in \mathcal{E}_D\}}{|\mathcal{E}_C|} & otherwise, \end{cases} \tag{4}$$

$$e\text{-}hd(\exists r.X, \exists s.Y) := \gamma(\nu + (1-\nu) \ .hd(\mathcal{T}_X, \mathcal{T}_Y)) \tag{5}$$

Definition 3.2 ($\mathcal{ELH}$ Similarity Degree)
The final similarity degree between concept C and D is defined by taking the average of homormorphism degree from C to D and D to C as follows:

$$\mathsf{sim}(C, D) = \frac{hd(\mathcal{T}_C, \mathcal{T}_D) + hd(\mathcal{T}_D, \mathcal{T}_C)}{2} \tag{6}$$

We can use the implementation of this measure on the website (http://ict.siit.tu.ac.th/sun.html). This measure also calculates the similarity based on two structural trees of concepts using the description tree algorithm. This measure is constructed using a specific language Description logic $\mathcal{ELH}$. But there are many primitive concepts in an ontology which their definitions are not sufficiently distinguish from each other. For this reason, ELSIM also cannot give the correct similarity degrees between all types of concepts in an ontolgy. Similarity degree between two concepts will be low if there has few inter-links between them but it will be high when ontology builder can add complete hierarchical information or more related links for these concepts. Another important thing is similarity values will be changed when we use different ontologies as ontologies have different structures. As a result, estimation of semantic similarity between ontology concepts based on taxonomic structure has the weakness for the concepts that have few hierarchical links in the ontology and we perform some experiments by applying existing two similarity measures mainly for the partially defined concepts or primitive concepts. Some evidence are shown in Table 1 and 2.

Table 1. Incomparable Similarity Values between Primitive Concepts using Path-based Measure and ELSIM with Human expert Results

Primitive Concept P_1	Primitive Concept P_2	Path-based	ELSIM	Human expert
Infiltrative lung tuberculosis	Nodular lung tuberculosis	0.2	0.0	0.7
maternal autoimmune hemolytic anemia	autoimmune hemolytic anemia	0.2	0.0	0.8
phakic corneal edema	Corneal epithelial edema	0.2	0.0	0.5

Table 2. Incomparable Similarity Values between primitive and defined concepts using Path-based Measure and ELSIM with Human expert Results

Primitive Concept P_1	Defined Concept P_2	Path-based	ELSIM	Human expert
Coronary artery thrombosis	Vertebral artery thrombosis	0.2	0.0	0.6
Corneal epithelial edema	Idiopathic corneal edema	0.1	0.0	0.6
Infectious mononucleosis hepatitis	chronic alcoholic hepatitis	0.2	0.0	0.5

3 Proposed Similarity Measure based on Concept Name

From the study of previous ontology-based similarity measures, they do not give the desired similarity degrees with the human expert result. If the ontology builders redefine the concepts with full relations, they will get higher similarity degree for path-based measures. Therefore, we want to fill the gap of ontology-based similarity measure and consider semantic similarity according to textual annotations (concept names) because each ontology concept is uniquely identified by a concept ID (e.g. id=10365005), annotated with a short textual description (e.g. "right main coronary artery thrombosis") and equipped with a definition in description logic. Moreover, ontology concept names are taken from the actual patient medical treatment records so they are very informative and can illustrate the complete meaning of the concept.

From this point of view, we propose a new similarity measure based on concept label from the natural language processing views. We modify concept name similarity by using following features.

1. Put different weights based on the headword of noun phrase to obtain a better similarity value
2. Use context-free grammar to compute the syntactic similarity based on the noun phrase structure of concept name.

3.1 Linguistic Headword Structure (Semantic Similarity)

All of the text labels of concept name are expressed in the form of noun phrase, in which the "headword" holds the core meaning of the phrase [14]. We cannot omit the headword in noun phrase. Therefore we should consider the highest weight for the headword when comparing the similarity of two concept names. In English, the structure of noun phrase can be defined as in the following cases.

1. Det + Pre-modifiers + noun (headword)
2. noun (headword) + Post-modifier/complement
3. noun + noun

All of the SNOMED CT concept names appear as the first case. Therefore, the rightmost noun is the headword of the concept name. After some experiments, we can conclude that the suitable weight for the headword is 0.6, and 0.4 is for the remaining components.

Let's consider concept P_1 = "right main coronary artery thrombosis" and concept P_2= "superior mesenteric vein thrombosis" For concept P_1,

- Weight for headword "thrombosis" is 0.6
- Weight for remaining components is 0.4 (0.1 for each remaining component)
- To assign different weights for each component, we consider positions of the component because the nearer component to the headword should get higher weight and it has higher semantic influence on the headword than other words [14]. Therefore, we give the weight for each component based on the distance from the headword. And then each component is divided by the distance value. For the component nearest from the headword, we subtract the sum of all other remaining components from 0.4. So, the sum of all weights of concept name is 1. As a result, the weight can be distributively estimated as shown in Table 3 and 4.

Table 3. Different weights of concept P_1

4 right	3 main	2 coronary	1 artery	0 thrombosis
0.1	0.1	0.1	0.1	0.6
0.1/4= 0.025	0.1/3= 0.033	0.1/2= 0.05	0.4-(0.025+0.033+0.05) = 0.292	0.6

Table 4. Different weights of concept P_2

3 superior	2 mesenteric	1 vein	0 thrombosis
0.133	0.133	0.133	0.6
0.133/3= 0.044	0.133/2= 0.067	0.4-(0.044+0.067)= 0.289	0.6

We define the headword noun pharse structure denoted by $\mathbf{sim}_{Headword}$ based on the Jaccard similarity [15] (the number of shared terms over the number of all unique tems). Therefore,

$|\text{wset}(P_1) \cap \text{wset}(P_2)|$ is the sum of the weights of shared terms and

$|\ \text{wset}(P_1) \cup \text{wset}(P_2)\ |$ is the sum of the weights of all unique terms.

$\mathbf{sim}_{Headword}(P_1,P_2)$

$$= \frac{|\ wset(P_1) \cap wset(P_2)\ |}{|\ wset(P_1) \cup wset(P_2)\ |}$$

$$= \frac{0.6}{(0.025 + 0.033 + 0.05 + 0.292 + 0.6 + 0.044 + 0.067 + 0.289)}$$

$$= 0.43$$

There are two points that we need to consider for this surface-matching similarity.

1. Some words are lexically similar but they have different meanings
   - For example, "kidney parenchyma" and "kidney beans"
   - "kidney parenchyma" is about human tissue of kidney and "kidney beans" is about a kind of bean.
   - In this case, it cannot occur because we compute the similarity based on the same category eg: for the disease category, all the concepts are about health such as illness, sickness and unwellness.
2. Some words are lexically different but they have similar meaning
   - For example, illness and sickness.
   - To fulfill this requirement, we used WordNet ontology to calculate the synsets similarity S_{synset} because two terms are similar if their synsets of these terms are lexically similar [16].

$$S_{synset}(h_1, h_2) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

     - A is the synset of headword h_1 and B is the synset of headword h_2
   - The main idea of our proposed method is based on the importance of two headwords terms. Therefore, we apply the synset similarity calculation to only the two important headwords. If the degree of similarity of synsets is greater 0, then the two words are considered to be the same. Otherwise, they are different.

$$Sim(h_1, h_2) = \begin{cases} 1, & \text{if } S_{synset}(h_1, h_2) > 0 \\ 0 & \text{if } S_{synset}(h_1, h_2) = 0 \end{cases} \tag{8}$$

3.2 Syntactic Structure Similarity

In order to know the syntactic structure of noun phrases for estimating the syntactic of the two noun phrases, we apply the context-free grammar (CFG) [17]. The grammar G= ⟨T, N, S, R⟩

- T is set of terminals
- N is set of non-terminals (NP in this case)
- S is the starting symbol
- R is rules or productions of the form

We construct noun phrase rules that cover all types of noun phrases in SNOMED CT concepts as listed in the following.

1. NP → N
2. NP → N NP
3. NP → Adj NP
4. NP → Det NP
5. NP → Adv NP

After applying CFG rule, the parsing orders of P_1 and P_2 are shown in the following list.

- Parsing order of P_1 : 3-3-3-2-1
- Parsing order of P_2 : 3-3-2-1

Syntactic similarity measure is estimated by the similarity of the applied CFG parsing rule. For the similarity calculation, numerator is the intersection of rules and denominator is the maximum number of rules.

$$\begin{aligned}\mathbf{sim}_{CFG}(P_1, P_2) &= \frac{4}{5} \\ &= 0.8\end{aligned}$$

3.3 Proposed Measure

After getting similarity values from two dimensions: headword structure and syntactic structure, we consider finalize similarity values by giving different weights based on their generalization. If two concepts are exactly same syntactic structure, but different headword terms, they have so much different meanings. But for the headword structure, it gives the accurate similarity value according to their headword position. This means that headword structure can decide the similarity more effective than syntactic structure. Accordingly, we make experiments by setting various weights as in Table 5.

Table 5. Different weights for headword and syntactic structure(CFG)

Concept P_1	Concept P_2	0.7 & 0.3	0.8 & 0.2	0.9 & 0.1	0.6 & 0.4	human result
Mosquito-borne hemorrhagic fever	glandular fever pharyngitis	0.5	0.4	0.3	0.6	0.5
Gangrenous paraesophageal hernia	Congenital bladder hernia	0.6	0.5	0.5	0.7	0.6

For the overall experiments, 0.7 and 0.3 get the highest correlation values. Therefore, we decide to set different weights as 0.7 for headword structure and 0.3 for syntactic structure.

$$\begin{aligned}\mathbf{Wsim}(P_1, P_2) &= a * \mathbf{sim}_{Headword}(P_1, P_2) + b * \mathbf{sim}_{CFG}(P_1, P_2) \\ &= 0.7 * 0.43 + 0.3 * 0.8 \\ &= 0.54\end{aligned}$$

4 Experimental Results on SNOMED CT

In the experiments, we use SNOMED CT which is the DL version released in January 2005 which contains 364,461 concept names. From the SNOMED CT

disorder category, 90 disease concept pairs are selected for evaluation of three cases so 30 concepts pairs for each experiment using path-based measure, ELSIM and our proposed measure. Similarity values between only primitive concepts are shown in Table 6. Similarity values between primitive and defined concepts are shown in Table 8 and Table 9 is shown the similarity results between only defined concepts. To examine the validity of all approaches, we evaluate the results of other two measures and our proposed method against human expert judgment. Five medical doctors make a consensus on the degree of similarity of the concepts as shown in each Table 6, 7 and 8.

4.1 Discussion

In order to evaluate the validaty of all measures against human expert result, we compute the correlation values and error values based on the results in Table 6, 7 and 8 and present in Table 9 and 10. As all of the ontology concepts are not completely structured with full relations and some concepts do not have enough hierarchical information in the ontology. Moreover, there is no useful hierarchical information for the primitive concepts. According to these facts, path-based and ElSIM offer very small correlation values for primitive concept similarity (0.04 and -0.19). ELSIM got the negative correlation so it means that these two results are totally different from each other. Based on the hard evidence, we assume to measure the similarity based on the concept name for the primitive concepts and our proposed method gets the highest correlation degree (0.74) and smallest error value (0.02) for the primitive concepts. Threrefore, it points out ontology concept names are also essential feature for the similarity measure between the concepts who do not have complete hierarchical information in the ontology. Concept names are taken from the actual patient medical treatment records, therefore, they are systematically constructed and very informative for each concept. Moreover, our proposed similarity measure calculates the similarity based on the linguistic headword structure by applying different weights and including Wordnet synsets similarity for headwords to include semantic similarity and also considers the similarity based on the syntactic structure. Therefore, our proposed measure gains benefit from both semantic and syntactic similarity of concept names.

In a consequence, our proposed method also gives a better accuracy (0.5) than other existing two approaches (0.2 and 0.18) for evaluating the similarity between primitive and defined concepts. For the case of defined concept, path-based and ELSIM get the highest correlation (0.61 and 0.6) when our proposed method gets small correlation (0.02) so it means ontology-based measures are the best similarity measure for the fully defined concepts in the ontology. But they cannot give the correct similarity degrees for the partially defined or primitive concepts who do not have the useful hierarchical information. But, our proposed measure overcomes the weakness of ontology-based measures for the primitive concepts. The most important merit of our proposed method is that it does not rely on the ontology structure, but can effectively capture the syntactic and semantic information of the concept names for the similarity measurement.

Table 6. Results of degree of similarity on 30 pairs between primitive concepts estimated by path-based method, ELSIM, our proposed method, and human expert

Primitive Concept P_1	**Primitive Concept** P_2	**Path-based**	**ELSIM**	**Proposed method**	**Human expert**
Hormonal tumor	Malignant mast cell tumor	0.2	0.0	0.5	0.6
Maternal autoimmune hemolytic anemia	Autoimmune hemolytic anemia	0.2	0.0	0.8	0.8
Hypertensive leg ulcer	Solitary anal ulcer	0.3	0.7	0.5	0.4
Bovine viral diarrhea	Bovine coronoviral diarrhea	0.6	0.6	0.7	0.7
Acute uterine inflammatory disease	Mycoplasmal pelvic inflammatory disease	0.4	0.2	0.9	0.9
Primary cutaneous blastomycosis	Primary pulmonary blastomycosis	0.7	0.9	0.7	0.6
Iodine-deficiency-related multinodular endemic goiter	Non-toxic multinodular goiter	0.8	0.7	0.8	0.8
Congenital pharyngeal polyp	Uterine cornual polyp	0.4	0.6	0.5	0.5
Phakic corneal edema	Corneal epithelial edema	0.2	0.0	0.5	0.5
Knee pyogenic arthritis	Gonococcal arthritis dermatitis syndrome	0.9	0.8	0.4	0.4
Hereditary canine spinal muscular atrophy	Spinal cord concussion	0.5	0.7	0.3	0.5
Mite-borne hemorrhagic fever	Meningococcal cerebrospinal fever	0.4	0.5	0.6	0.5
Congenital cleft larynx	Congenital spastic foot	0.6	0.8	0.3	0.3
Congenital acetabular dysplasia	Short rib dysplasia	0.5	0.9	0.5	0.5
Intestinal polyposis syndrome	Ovarian vein syndrome	0.6	0.8	0.6	0.5
Extrapulmonary subpleural pulmonary sequestration	Pulmonary alveolar proteinosis	0.7	0.6	0.4	0.4
Atypical chest pain	Psychogenic back pain	0.3	0.1	0.5	0.5
Puerperal pelvic cellulitis	Chronic female pelvic cellulitis	0.9	0.7	0.8	0.7
Spinal cord hypoplasia	Spinal cord rupture	0.5	0.7	0.6	0.6
Infiltrative lung tuberculosis	Nodular lung tuberculosis	0.2	0.0	0.9	0.7
Early gastric cancer	Primary vulval cancer	0.4	0.8	0.4	0.4
Congenital mesocolic hernia	Gangrenous epigastric hernia	0.2	0.0	0.5	0.4
Congenital nonspherocytic hemolytic anemia	Congenital macular corneal dystrophy	0.2	0.0	0.3	0.2
Congenital cerebellar cortical atrophy	Congenital renal atrophy	0.6	0.9	0.7	0.2
Puerperal pyrexia	Heat pyrexia	0.3	0.0	0.5	0.6
Methylmalonyl-CoA mutase deficiency	Muscle phosphoglycerate mutase deficiency	0.2	0.0	0.7	0.5
Recurrent mouth ulcers	Multiple gastric ulcers	0.5	0.8	0.4	0.4
Infantile breast hypertrophy	Sebaceous gland hypertrophy	0.6	0.7	0.6	0.4
Congenital pyloric hypertrophy	Synovial hypertrophy	0.3	0.0	0.5	0.3
Inflammatory testicular mass	Inflammatory epidermal nevus	0.5	0.7	0.3	0.3

Table 7. Results of degree of similarity on 30 pairs between primitive and defined concepts estimated by path-based method, ELSIM, our proposed method, and human expert

Primitive Concept P_1	**Defined Concept** P_2	**Path-based**	**ELSIM**	**Proposed method**	**Human expert**
Mosquito-borne hemorrhagic fever	Glandular fever pharyngitis	0.4	0.7	0.5	0.5
Right main coronary artery thrombosis	Coronary artery rupture	0.9	0.9	0.5	0.4
right main coronary artery thrombosis	superior mesenteric vein thrombosis	0.7	0.9	0.5	0.6
Infectious mononucleosis hepatitis	chronic alcoholic hepatitis	0.2	0.0	0.5	0.5
Cerebral venous sinus thrombosis	Phlebitis cavernous sinus	1.0	0.9	0.6	0.6
Third degree perineal laceration	Complex periorbital laceration	0.3	0.7	0.5	0.5
Congenital subaortic stenosis	Rheumatic aortic stenosis	0.9	0.7	0.6	0.7
Congenital acetabular dysplasia	Aortic valve dysplasia	0.5	0.6	0.5	0.3
Intestinal polyposis syndrome	Fetal cytomegalovirus syndrome	0.4	0.4	0.6	0.3
Anterior choroidal artery syndrome	Juvenile polyposis syndrome	0.4	0.7	0.5	0.3
Puerperal pelvic cellulitis	Streptococcal cellulitis	0.3	0.5	0.5	0.3
Benign hypertensive renal disease	Pulmonary hypertensive venous disease	0.7	0.8	0.6	0.4
Corneal epithelial edema	Idiopathic corneal edema	0.1	0.0	0.8	0.6
Chronic sarcoid myopathy	Hereditary hollow viscus myopathy	0.3	0.6	0.5	0.5
Primary cutaneous blastomycosis	Chronic pulmonary blastomycosis	0.7	0.9	0.6	0.6
Gingival pregnancy tumor	Granular cell tumor	0.4	0.5	0.6	0.4
Borderline epithelial tumor	Melanotic malignant nerve sheath tumor	0.4	0.6	0.4	0.4
Congenital sternomastoid tumor	Malignant mast cell tumor	0.4	0.5	0.4	0.4
Congenital pharyngeal polyp	Rhinosporidial mucosal polyp	0.4	0.4	0.6	0.5
Mercurial diuretic poisoning	Lobelia species poisoning	0.4	0.4	0.4	0.5
Branch macular artery occlusion	Acute mesenteric arterial occlusion	0.5	0.9	0.5	0.6
Intrarenal hematoma	Stomach hematoma	0.5	0.9	0.5	0.6
Spinal cord hypoplasia	Spinal cord dysplasia	0.9	0.9	0.6	0.6
Coronary artery thrombosis	Vertebral artery thrombosis	0.2	0.0	0.9	0.6
Duodenal papillary stenosis	Congenital bronchial stenosis	0.5	0.6	0.6	0.4
Arteriovenous fistula stenosis	Subclavian vein stenosis	0.4	0.2	0.6	0.5
Mechanical hemolytic anemia	Hereditary sideroblastic anemia	0.7	0.7	0.6	0.5
Malignant catarrhal fever	Malignant lipomatous tumor	0.7	0.2	0.3	0.3
Bolivian hemorrhagic fever	Dengue hemorrhagic fever	0.6	0.9	0.8	0.6
Benign brain tumor	Benign neuroendocrine tumor	0.4	0.0	0.5	0.5

Table 8. Results of degree of similarity on 30 Defined disease concepts estimated by path-based method, ELSIM, our proposed method, and human expert

Defined Concept P_1	**Defined Concept** P_2	**Path-based**	**ELSIM**	**Proposed method**	**Human expert**
Rheumatic heart valve stenosis	Coronary artery stenosis	0.6	0.8	0.5	0.5
Nasal septal hematoma	Vocal cord hematoma	0.3	0.9	0.5	0.8
Simple periorbital laceration	Brain stem laceration	0.5	0.9	0.4	0.8
Peritonsillar cellulitis	Dentoalveolar cellulitis	0.5	0.9	0.6	0.6
Parainfluenza virus laryngotracheitis	Acute viral laryngotracheitis	1.0	0.9	0.4	0.9
Bone marrow hyperplasia	Retromolar gingival hyperplasia	0.8	0.8	0.5	0.8
Chronic proctocolitis	Chronic viral hepatitis	0.5	0.8	0.4	0.5
Obstructive biliary cirrhosis	Syphilitic portal cirrhosis	0.6	0.8	0.5	0.6
Peripheral T-cell lymphoma	Primary cerebral lymphoma	0.5	0.8	0.5	0.6
Mast cell leukemia	Prolymphocytic leukemia	0.9	0.9	0.4	0.9
Tricuspid valve regurgitation	Rheumatic mitral regurgitation	0.9	0.9	0.5	0.8
Gangrenous paraesophageal hernia	Congenital bladder hernia	0.5	0.8	0.6	0.6
Congenital mandibular hyperplasia	Atypical endometrial hyperplasia	0.3	0.6	0.5	0.7
Tuberculous adenitis	Acute mesenteric adenitis	0.7	0.6	0.5	0.6
Congenital skeletal dysplasia	Aortic valve dysplasia	0.4	0.7	0.5	0.7
Histiocytic sarcoma	Alveolar soft part sarcoma	1.0	0.9	0.5	0.8
Drug-induced ulceration	Amebic perianal ulceration	0.9	0.6	0.5	0.7
Cervical radiculitis	Cervical lymphadenitis	0.4	0.8	0.7	0.7
Basilar artery embolism	Obstetric pulmonary embolism	0.6	0.7	0.5	0.6
Acute apical abscess	Chronic apical abscess	0.5	1.0	0.9	0.8
Acute glossitis	Chronic glossitis	0.4	0.4	0.5	0.6
Acute bronchitis	Acute purulent meningitis	0.3	0.6	0.4	0.4
Acute lower gastrointestinal hemorrhage	Stromal corneal hemorrhage	0.4	0.8	0.4	0.4
Epidural hemorrhage	Tracheostomy hemorrhage	0.6	0.7	0.5	0.6
Thallium sulfate toxicity	Ammonium sulfamate toxicity	1.0	0.6	0.6	0.6
Simple periorbital laceration	Complex periorbital laceration	0.8	1.0	0.8	0.9
Biceps femoris tendinitis	Profunda femoris artery thrombosis	0.5	0.8	0.2	0.6
Hyperplastic thrush	Hyperplastic gingivitis	0.3	0.5	0.7	0.5
Acer rubrum poisoning	Penicillium rubrum toxicosis	0.5	0.6	0.6	0.6
Acute vesicular dermatitis	Herpesviral vesicular dermatitis	0.4	0.5	0.9	0.4

Table 9. Correlation Values between Similarity Measures and human experts for each case

Method	Type	primitive concepts	primitive and defined	defined concepts
Path-based	ontology-based	0.04	0.2	0.61
ELSIM	ontology-based	-0.19	0.18	0.6
Proposed method	concept name	0.74	0.5	0.02

Table 10. Error Values between Similarity Measures and human experts for each case

Method	Type	Primitive concepts	Primitive and Defined	Defined concepts
Path-based	ontology-based	0.1	0.1	0.04
ELSIM	ontology-based	0.2	0.1	0.03
Proposed method	concept name	0.02	0.02	0.06

5 Conclusions

This research reviews the existing ontology-based semantic similarity measures and points out the limitations of these measures based on the three cases of experiment. To overcome the weakness, a new similarity measure has been proposed based on the similarity measure of the semantic and syntactic structure of concept name. This paper also shows the strength and weakness of different measures (based on hierarchical information and concept names) according to the evaluation results against human experts.

This research has many continuous works. Firstly, we are planning to combine our proposed method and existing ontology-based measures to fill the weakness of each other so we will get the perfect and desired similarity results for all different types of concepts. As the next step, we will evaluate the new combined similarity method against more number of different concepts. Finally, we want to apply the new combined similarity method to other medical ontologies such as MeSH ontology, to evaluate the effectiveness of our proposed method.

Acknowledgments

The first author is supported by Sirindhorn International Institute of Technology (SIIT) under an Excellent Foreign Student (EFS) scholarship. We are thankful to all those anonymous experts who spent their valuable time to evaluate the similarity between diseases.

References

1. P. Resnik. Semantic Similarity in a Taxonomy. An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, pp. 95-130, 1999.
2. A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G.M. Petrakis and E. Milios: Information retrieval by Semantic Similarity. International Journal of Semantic Web Informatics Systems, pp. 55-73, 2006.
3. T. Pedersen, S. Pakhomov, S. Patwardhan and C. Chute: Measures of Semantic Similarity and Relatedness in the Biomedical Domain: Journal of Biomedical Informatics, pp. 288-299, 2007.
4. S. Tongphu, B. Suntisrivaraporn, B. Uyyanonvara and M. N. Dailey: Ontology-based Object Recognition of Car Sides. In Proceedings of 9th International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI), 2012.
5. SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terminology. http://www.snomed.org/snomedct/index.html.
6. V. N. Garla and C. Brandt: Semantic Similarity in the Biomedical Domain: An Evaluation across Knowledge Sources. Journal of BMC Bioinformatics October 2012.
7. A. Abdelrahman and A. Kayed: A Survey on Semantic Similarity Measures between Concepts in Health Domain: American Journal of Computational Mathematics, pp. 204-214, 2015.
8. M. Zare, C. Pahl, M. Nilashi, N. Salim and O. Ibrahim: A Review of Semantic Similarity Measures in Biomedical Domain Using SNOMED CT: Journal of Soft Computing and Decision Support Systems, Vol.2, No.6, September 2015.
9. I. Choi and M. Kim: Topic Distillation using Hierarchy Concept Tree. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada; pp. 371-72, 2003.
10. H. AI-Mubaid and H. Nguyen: A Cluster-based Approach for Semantic Similarity in the Biomedical Domain. In Proceedings of the 28th IEEE EMBS Annual International Conference. New York City, USA, Sept 3, 2006.
11. M. Batet, D. Sanchez and A. Valls: An Ontology-based Measure to Compute Semantic Similarity in Biomedicine: Journal of Biomedical Informatics, pp. 118-125, 2011.
12. Nowlan and Kay: SNOMED CT Basics, IHTSDO, International Health Terminology Standards Development Organization, August 2008.
13. S. Tongphu and B. Suntisrivaraporn: Algorithms for Measuring Similarity Between ELH Concept Descriptions: A Case Study on SNOMED CT. Journal of Computing and Informatics, Vol-20, Jul-8, 2015.
14. M. Liberman and R. Sproat: The Stress and Structure of Modified Noun Phrases in English, Stanford University, 1992.
15. W. H. Gomaa and Aly A. Fahmy: A Survey of Text Similarity Approaches: International Journal of Computer Applications, Vol-68, No.13, April 2013.
16. E. G. M. Petrakis, G. V. A. Hliaoutakis, P. Raftopoulou: X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies: Journal of Digital Information Management, April 2006.
17. S. Ko, Y. Han and K. Salomma: Approximate Matching between a Context-free grammar and a Finite-state Automaton: Information and Computation: pp. 278-289, February, 2016.