Text Similarity Approach for SNOMED CT Primitive Concept Similarity Measure

Htet Htet Htun
School of Information, Computer and Communication Technology,
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
Email: htethtethtun.8910@gmail.com

Virach Sornlertlamvanich
School of Information, Computer and Communication Technology,
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
Email: virach@siit.tu.ac.th

Abstract—**For the biomedical ontologies, Concept Similarity Measures (CSMs) become important in order to find similar treatments between diseases. For the ontology primitive concepts, they do not have enough definitions because they are partially defined in the ontology so one way to find the similarity between primitive concepts is to apply textual similarity methods between concept names. But existing textual similarity methods cannot give correct similarity degrees for all concept pairs. In this paper, we propose a new primitive concept name similarity measure based on natural language processing to get a better result in concept similarity measure in terms of noun phrase construction analysis. We conduct experiments on the standard clinical ontology SNOMED CT and make the comparison between our proposed method and existing two approaches against human expert results in order to prove our proposed similarity measure give correct and nearest similarity degree between primitive concepts.**

Keywords — **Primitive Concept Similarity Measure, Text Similarity, Natural Language Processing, SNOMED CT, Description Logic**

I. Introduction

Semantic similarity measures are widely used in Natural Language Processing and existing similarity measures have been adapted to the biomedical domain. In Description Logic (DL), concept similarity measure (CSM) also has been proposed for the biomedical domain. It determines the similarity between two concepts and returns the numerical value between 0 and 1 that represents their similarity degree [9]. It is the main task especially for the case of finding similar treatments between two diseases. For a health decision support system, if we know the similarity between two diseases, we can find and conclude the similar treatments for those diseases based on their level of similarity. Therefore, knowing the level of similarity between two diseases is the important task for the biomedical ontologies such as SNOMED CT which is a standard terminology that covers all areas of clinical information including body structure, diseases, organisms and clinical finding etc. In the SNOMED CT that is written in DL, there are two kinds of concepts - defined concepts and primitive concepts as the following.

Hypoxia of brain $\equiv$ Hypoxia $\sqcap$ $\exists$FindingSite. Brain Structure
Tumor of dermis $\sqsubseteq$ Navigational concept
Vibrio species n-z $\sqsubseteq$ Navigational concept

In the above three concepts, “hypoxia of brain” is the defined concept because it has “is-a” relationship with “hypoxia” and “attribute-value” relationship type “findingSite” with another concept “brain structure”. So, its definition is sufficient to distinguish from all other concepts’ definitions [2]. But another two concepts, “tumor of dermis” and “vibrio species n-z”, they have only “is-a” relationship such as “tumor of dermis” is a navigational concept and “vibrio species n-z” is a navigational concept. Therfore, their definitions are the same and not sufficient to distinguish from each other. It means primitive concepts do not have enough definitions and their definitions are needed to define with additional information [6]. So, we cannot find the similarity of primitive concepts based on their definitions. For this reason, we find the primitive concepts similarity based on the textual annotations (concept names) using text similarity methods. But existing text similarity methods give incorrect and unsuitable similarity values for some concept pairs [1] because they are surface-matching similarity measures and they treat the same weight for all positions in concept names like Jaccard similarity [10]. Therefore, we propose a new concept name similarity measure based on two different similarities in order to get nearest similarity values as the human expert results and then we intend to find the similar treatments based on their similarity values between diseases.

The rest of the paper is organized in the following order. Section II reviews the background of concept similarity measure in DL and SNOMED CT that we apply for the experiments. Section III presents our proposed similarity measure and calculation based on concepts in Table I. Section IV explains the evaluation of proposed method and correlation values with human results. Finally, section V presents the conclusion and future work.

II. Preliminaries

In Description Logics (DLs), concept descriptions are inductively defined with a set of concept names CN and a set of role names RN. The set of concept descriptions for a specific

DL $\mathcal{ELH}$ is denoted by Con($\mathcal{ELH}$) [4]. The set Con($\mathcal{ELH}$) can be defined as follow:

$$\mathsf{C}, \mathsf{D} \rightarrow \mathsf{A} \mid \mathsf{T} \mid \mathsf{C} \sqcap \mathsf{D} \mid \exists r.\mathsf{C}$$

where T denotes the top concept, C, $\mathsf{D} \in$ Con($\mathcal{ELH}$), A is concept name and r is role name. In DL, concept names appearing on the left-hand side of a definition are called defined concept names (CN^{def}). Other concept names are called primitive concept names (CN^{pri}) [4]. Therefore, CN= $\mathsf{CN}^{pri} \cup \mathsf{CN}^{def}$. In Description Logic, primitive concept names similarity is defined as the following.

Primitive concept similarity : Let $\mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$ be a set of primitive concept names occurring in terminology $\mathcal{T}$. A primitive concept similarity is a partial function [1] s^c: $\mathsf{CN} \times \mathsf{CN} \rightarrow$ [0,1], where $\mathsf{CN} \subseteq \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$. For two concept names $\mathsf{A},\mathsf{B} \in \mathsf{CN}^{\mathsf{pri}}(\mathcal{T})$, s^c (A,B) = s^c (B,A) and s^c(A,A) =1.

Let $\mathsf{CN}^{\mathsf{pri}}$ be the set of all primitive concept names in SNOMED CT. For each $P \in \mathsf{CN}^{\mathsf{pri}}$, we denote by text(P), the textual annotation (concept name) of P. For convenience, we denote by $tset(P)$, the set of words occurring in $text(P)$ as Table I.

TABLE I
SNOMED CT PRIMITIVE CONCEPTS

Primitive concepts P	conceptID (P)	text(P)	tset(P)
P_1	19036004	rheumatic heart valve stenosis	{"rheumatic","heart","valve", "stenosis" }
P_2	233970002	coronary artery stenosis	{"coronary","artery", "stenosis" }

For primitive concept similarity, we used the concept names because ontology concept names are taken from the actual patient medical treatment records so they are very informative and can illustrate the complete meaning for the concept.

III. PROPOSED PRIMITIVE CONCEPT NAME SIMILARITY MEASURE

In SNOMED CT ontology, all concept names are expressed in the form of noun phrase (NP). In NP, they have the "headword" that holds the core meaning of the 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 concept name. Moreover, each component in noun phrase should get different weights based on their positions. Therefore, we modified the similarity according to their linguistic structure by using following two different similarities.

1) Put different weights to each component based on the headword of noun phrase to obtain a better similarity value and
2) Compute the syntactic similarity based on the noun phrase structure of concept name using context-free grammar.

A. Linguistic Headword Structure (Semantic Similarity)

After some experiments, we can conclude that the suitable weight for the headword is 0.6 and 0.4 is for the remaining components.

For concept P_1 in Table I,

- Weight for headword "stenosis" is 0.6
- Weight for remaining components is 0.4 (0.133 for each remaining component)
- To give different weights for each component, we considered 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 [7]. Therefore, we considered 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 II and III.

TABLE II
DIFFERENT WEIGHTS OF CONCEPT P_1

3 rheumatic	2 heart	1 valve	0 stenosis
0.133	0.133	0.133	0.6
0.133/3= 0.044	0.133/2= 0.066	0.4- (0.044+0.066) = 0.29	0.6

TABLE III
DIFFERENT WEIGHTS OF CONCEPT P_2

2 coronary	1 artery	0 stenosis
0.2	0.2	0.6
0.2/2= 0.1	0.4-(0.1)= 0.3	0.6

We use the Jaccard method for the similarity of headword structure denoted by $\mathbf{sim}_{Headword}$.

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

$$= \frac{\mid tset(P_1) \cap tset(P_2) \mid}{\mid tset(P_1) \cup tset(P_2) \mid}$$
$$= \frac{0.6}{(0.044 + 0.066 + 0.29 + 0.6 + 0.1 + 0.3)}$$
$$= 0.43$$

There are two points have to consider for the surface-matching similarity.

1) Some words are lexically similar but have different meanings
   - E.g., “kidney parenchyma” and “kidney beans”
   - The “kidney parenchyma” is about human tissue of kidney and “kidney beans” is about one kinds of beans.
   - 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 have similar meaning
   - E.g., illness and sickness. They have very similar meaning but different terms.
   - 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 [8].

$$S_{synset}(P_1, P_2) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

     – A is the synset of concept P_1 and B is the synset of concept P_2
   - Therefore, we apply the synset similarity only for important two headwords. If similarity degree of synsets greater than 0, considered those two words are the same. Otherwise, those two words are different.

$$Sim(P_1, P_2) = \begin{cases} 1, & \text{if } S_{synset}(P_1, P_2) > 0 \\ 0 & \text{if } S_{synset}(P_1, P_2) = 0 \end{cases} \quad (2)$$

B. 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) [5]. The grammar is 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

For two concepts in Table I,

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

For the similarity calculation, nominator is the intersection of rules and denominator is the maximum number of rules.

$$\begin{aligned} \mathbf{sim}_{CFG}(P_1, P_2) &= \frac{3}{4} \\ &= 0.75 \end{aligned}$$

C. Proposed Similarity 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. Therefore, we decide to set different weights as 0.7 for headword structure and 0.3 for syntactic structure.

$$\begin{aligned} \mathbf{Wsim}(P_1, P_2) &= a * \mathbf{sim}_{Headword}(P_1, P_2) + b * \mathbf{sim}_{CFG}(P_1, P_2) \\ &= 0.7 * 0.43 + 0.3 * 0.75 \\ &= 0.53 \end{aligned}$$

IV. Experimental Results on SNOMED CT

For the experiments, we use SNOMED CT which is the DL version released in January 2005 which contains 364,461 concept names. In SNOMED CT, each concept name is uniquely identified by a concept ID (eg: id=19036004), annotated with a short textual description (eg: “rheumatic heart valve stenosis”). From the disorder category, we evaluate 30 disease concepts using our proposed measure as shown in Table IV. The usual way of evaluating the accuracy of proposed measure is to compare with human expert results. So, we request similarity values of 30 concept pairs from 3 medical doctors who have already worked in the hospital and 2 medical doctors graduated in a last few months. After discussing and making normalization, final human expert results are shown as in Table IV.

A. Discussion

Our Proposed similarity measure calculate the similarity based on the linguistic headword structure by applying different weights and including Wordnet synonym sets similarity for headwords to include semantic similarity. Moreover, our proposal also considers the similarity based on the syntactic structure. Therefore, our proposed measure gains benefit from both semantic and syntactic similarity of concept names. But if we use the simple Jaccard measure for two concepts in Table I,

$$\mathbf{sim}^{Jaccard}\ (P_1,P_2)= \frac{|tset(P_1) \cap tset(P_2)|}{|tset(P_1) \cup tset(P_2)|} = \frac{1}{6} = 0.167$$

we will get very few similarity value. But human expert give the similarity value for two concepts in Table I as 0.6.

TABLE IV
RESULTS OF DEGREE OF SIMILARITY ON 30 DISEASE CONCEPTS ESTIMATED BY OUR PROPOSED METHOD AND HUMAN EXPERT

Concept P_1	Concept P_2	Proposed method	Human expert result
Hormonal tumor	Malignant mast cell tumor	0.5	0.6
Maternal autoimmune hemolytic anemia	Autoimmune hemolytic anemia	0.8	0.8
Hypertensive leg ulcer	Solitary anal ulcer	0.5	0.4
Bovine viral diarrhea	Bovine coronoviral diarrhea	0.7	0.7
Liver cell carcinoma	Blastomycosis liver	0.7	0.6
Right main coronary artery thrombosis	Superior mesenteric vein thrombosis	0.5	0.6
Acute uterine inflammatory disease	Mycoplasmal pelvic inflammatory disease	0.9	0.9
Primary cutaneous blastomycosis	Primary pulmonary blastomycosis	0.7	0.6
Corneal ulcer	Acute gastrojejunal ulcer	0.4	0.6
Iodine-deficiency-related multinodular endemic goiter	Non-toxic multinodular goiter	0.8	0.8
Cerebral venous sinus thrombosis	Phlebitis cavernous sinus	0.6	0.6
Complex periorbital laceration	Third degree perineal laceration	0.5	0.5
Congenital pharyngeal polyp	Uterine cornual polyp	0.5	0.5
Mosquito-borne hemorrhagic fever	Glandular fever pharyngitis	0.5	0.5
Phakic corneal edema	Corneal epithelial edema	0.5	0.5
Rheumatic heart valve stenosis	Coronary artery stenosis	0.5	0.6
Knee pyogenic arthritis	Gonococcal arthritis dermatitis syndrome	0.4	0.4
Hereditary canine spinal muscular atrophy	Spinal cord concussion	0.3	0.5
Simple periorbital laceration	Brain stem laceration	0.4	0.5
Intestinal polyposis syndrome	Ovarian vein syndrome	0.6	0.5
Mite-borne hemorrhagic fever	Meningococcal cerebrospinal fever	0.6	0.5
Nasal septal hematoma	Vocal cord hematoma	0.5	0.5
Congenital cleft larynx	Congenital spastic foot	0.3	0.3
Congenital acetabular dysplasia	Short rib dysplasia	0.5	0.5
Extrapulmonary subpleural pulmonary sequestration	Pulmonary alveolar proteinosis	0.4	0.4
Congenital subaortic stenosis	Rheumatic aortic stenosis	0.6	0.7
Infectious mononucleosis hepatitis	Chronic alcoholic hepatitis	0.5	0.5
Coronary artery rupture	Right main coronary artery thrombosis	0.5	0.4
Atypical chest pain	Psychogenic back pain	0.5	0.5
Puerperal pelvic cellulitis	Chronic female pelvic cellulitis	0.8	0.7

So, it didn't match with the human expert results. According to our proposed method, similarity value is 0.53 so it makes high improvement than the existing surface-matching text similarity measures. In order to evaluate our proposed measure against human expert results, we compute the correlation values for our accuracy based on the results in Table IV. We got 0.83 correlation values with human expert results as shown in Table V. Therefore, our proposed similarity measure is essential for primitive concepts measure and can give the nearest similarity degrees as the human results.

TABLE V
CORRELATION VALUES BETWEEN PROPOSED MEASURE AND HUMAN EXPERTS

Method	Method Type	Correlation
Proposed measure	textual annotations/ concept names	0.83

After getting the correlation value, we wish to know overall difference or error between our proposed method and human expert results. Therefore, we calculate "mean squared error" (MSE) which measures the average of the squares of "errors" between two results. We obtain 0.0067 which is very close to zero so it means our proposed method gives the same results as the human experts. To measure the quality of our proposed method, we calculate not only the correlation value but also the error between our proposed method and human expert results.

In order to prove our novelty and performance completely, we compare our proposed method to the existing two similarity methods (path-based and ELSIM). The first method finds the similarity between two concepts based on the taxonomic structure of the ontology [3]. It considers all of the superconcepts belonging to all the possible taxonomical 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 c1 and c2 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. The second method is ELSIM that computes the similarity between ELH concepts based on homomorphism tree function. This measure is used for description logic ELH definitions and provides a numerical value that represents structural similarity of one concept description against to the another concept description (see in detail [4]). The implementation of this method is on the website (http://ict.siit.tu.ac.th/sun.html). From the Table IV, we pick up the first 20 pairs to compute the similarity degrees by applying path-based, ELSIM, proposed measure and human expert as shown in Table VI.

And then, we calculate the correlation values of all methods against the human result as shown in Table VII. According to the correlation results, the existing two measures get the very small correlation values (0.005 and -0.38). ELSIM gets the negative correlation so it means that these two results are totally different from each other. This shows that these two measures cannot give the accurate similarity degrees between primitive concepts because they have limited information and actually need to redefine with complete information. As a result, our proposed measure gets the highest correlation value (0.82) among existing two measures. Therefore, we can conclude that our proposed measure is essential and gives very accurate results for the primitive concept similarity measure.

TABLE VI

RESULTS OF DEGREE OF SIMILARITY ON 20 PAIRS BETWEEN PRIMITIVE CONCEPTS ESTIMATED BY PATH-BASED METHOD, ELSIM, OUR PROPOSED METHOD, AND HUMAN EXPERT

Concept P_1	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
Liver cell carcinoma	Blastomycosis liver	0.4	0.4	0.7	0.6
Right main coronary artery thrombosis	Superior mesenteric vein thrombosis	0.7	0.9	0.5	0.6
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
Corneal ulcer	Acute gastrojejunal ulcer	0.5	0.8	0.4	0.6
Iodine-deficiency-related multinodular endemic goiter	Non-toxic multinodular goiter	0.8	0.7	0.8	0.8
Cerebral venous sinus thrombosis	Phlebitis cavernous sinus	1.0	0.9	0.6	0.6
Complex periorbital laceration	Third degree perineal laceration	0.3	0.7	0.5	0.5
Congenital pharyngeal polyp	Uterine cornual polyp	0.4	0.6	0.5	0.5
Mosquito-borne hemorrhagic fever	Glandular fever pharyngitis	0.4	0.7	0.5	0.5
Phakic corneal edema	Corneal epithelial edema	0.2	0.0	0.5	0.5
Rheumatic heart valve stenosis	Coronary artery stenosis	0.6	0.8	0.5	0.6
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
Simple periorbital laceration	Brain stem laceration	0.5	0.9	0.4	0.5
Intestinal polyposis syndrome	Ovarian vein syndrome	0.6	0.8	0.6	0.5

TABLE VII

CORRELATION VALUES BETWEEN PATH-BASED, ELSIM, PROPOSED MEASURE AND HUMAN EXPERTS

Method	Method Type	Correlation
Path-based measure	ontology-based	0.005
ELSIM	ontology-based	-0.38
Proposed measure	textual annotations/ concept names	0.82

V. CONCLUSION AND FUTURE WORK

This paper proposed a new primitive concept name similarity measure based on semantic and syntactic similarities. We prove the accuracy of proposed method by calculating correlation value and error value against human expert results. Moreover, we compare our proposed method to the existing two approaches in order to prove that our proposed method outperforms the existing approaches.

There are some directions for our future work. Firstly, we will evaluate our approach with more number of primitive concepts. Secondly, we will compare the accuracy of our approach with other existing ontology-based approaches for ontology primitive concepts similarity. Thirdly, we will evaluate our proposed measure for the defined concepts to confirm the validity of our approach on both two kinds of concepts - defined concepts and primitive concepts. Finally, we will analyze benefits and usability of our approach and other similarity approaches for both defined concepts and primitive concepts.

ACKNOWLEDGMENT

The first author is supported by Sirindhorn International Institute of Technology (SIIT) under an Excellent Foreign Student (EFS) scholarship.

REFERENCES

[1] H.H.Htun, V.Sornlertlamvanich and B.Suntisrivaraporn: Towards Automatic Generation of Preference Profile for Primitive Concept Similarity Measures on SNOMED CT. In Proceedings of the 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), pp. 194-199, Yogyakarta, Indonesia, 2016.

[2] Nowlan and Kay: SNOMED CT Basics, IHTSDO, International Health Terminology Standards Development Organization, August 2008.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M.Liberman and R.Sproat: The Stress and Structure of Modified Noun Phrases in English, Stanford University, 1992.

[8] 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.

[9] T.Racharak, B.Suntisrivaraporn and S.Tojo. sim^{π}: A Concept Similarity Measure under an Agents Preferences in Description Logic $\mathcal{ELH}$. In Proceedings of the 8th the International Conference on Agents and Artificial Intelligence (ICAART 2016). Rome, Italy, 2016.

[10] W.H. Gomaa and Aly A. Fahmy: A Survey of Text Similarity Approaches: International Journal of Computer Applications, Vol-68, No.13, April 2013.