September 2016

SNOMED CT Primitive Concept Similarity Measure by Concept Name Text Similarity Approach

Htet Htet Htun [1] and Virach Sornlertlamvanich
School of Information, Computer and Communication Technology,
Sirindhorn International Institute of Technology, Thammasat University, Thailand

Abstract. In the last few years, Concept Similarity Measures (CSMs) become important for the biomedical ontologies in order to find adaptable treatments from the conceptually similar diseases. For the ontology primitive concepts, they are not fully defined in the ontology so taxonomical path-based similarity measure cannot give the correct similarity for primitive concepts. 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 comparison between taxonomical path-based measure and our proposed similarity measure against human expert results in order to prove our proposed similarity measure can outperform the existing approaches for primitive concept similarity.

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

1. Introduction

To determine the similarity between word pairs is an important work in text understanding and it has been applied in many natural language processing tasks like the processing, classification, clustering and structuring of textual resources. In Description Logic (DL), concept similarity measure (CSM) 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 degree of similarity [8]. It is the main task especially for a health decision support system retrieving similar treatment cases from the past hospital database as guidelines to treat the current patient. For example, the similarity of two diseases "thoracic nerve root pain" and "cervical nerve root pain" is 85%, the system gives the same or similar treatment (Gabapentin) for these two diseases. Therefore, knowing similarity levels between diseases is beneficial to find the alternative treatments for those diseases. It actually assists to find related treatments or solutions for the current patient based on previous recorded diseases. Therefore, concept similarity measure becomes the main factor in the biomedical domain and many researchers focus semantic similarity ap-

[1]Corresponding Author: Email: htethtethtun.8910@gmail.com

proaches based on the biomedical ontologies such as SNOMED CT or MeSH. SNOMED CT 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

- Is-a = hypoxia
- Finding site = Brain structure
- Sufficiently Defined (defined concept)

Tumor of dermis

- Is-a = navigational concept
- Primitive concept

Vibrio species n-z

- Is-a = navigational concept
- Primitive concept

In the above three concepts, "hypoxia of brain" is the defined concept because it has "is-a" relation with "hypoxia" and "attribute-value" relation type "finding site" with another concept "brain structure". So, its definition is sufficient to distinguish from all other concepts' definitions [3]. But another two concepts, "tumor of dermis" and "vibrio species n-z", they have only "is-a" relation 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 that primitive concepts do not have enough definitions and they need to be defined with additional information. For this reason, we cannot find correct similarity values between primitive concepts based on their original definitions. Instead, we find the primitive concept similarity based on the text label of the concept. And 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 2 presents related work of concept similarity measures. Section 3 reviews ontology taxonomical-based similarity measure. Section 4 presents our proposed similarity measure. Section 5 explains the evaluation of ontology path-based measure, proposed method and correlation values with human results based on SNOMED CT ontology. Finally, section 6 presents the conclusion and future work.

2. Related Work

There have been proposed many approaches for the semantic similarity measure between concepts by exploiting medical ontologies (SNOMED CT or MeSH). But most of existing approaches measure the similarity based on the ontology structure as the primary source by taking the minimum number of shortest path between evaluated two concepts. Wu and Palmer [1] and [2] proposed a path-based measure that also takes into account

the depth of the concepts in the hierarchy and Choi and Kim [4] also proposed according to the difference on the depth levels of two concepts and the distance of the shortest path between them. In [2], they proposed a new path-based measure based on SNOMED CT ontology as the experiment and they showed that their method gets the highest accuracy among other path-based measures. So, we review this method in section 3 and point out the weakness of path-based measure. After that, we proposed primitive concept similarity measure based on the concept names.

3. Taxonomical Path-based Measure

This measure computes the similarity based on the taxonomical paths connecting the two concepts [2]. 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 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) = $\{\ c_j \in \mathrm{C} \mid c_j$ is superconcept of $c_i\ \} \cup \{\ c_i\ \}$ is defined as the union of the ancestors of the concept c_i and c_i itself.

4. Proposed Similarity Measure based on Concept Name

The previous measure (in section 3) computes the similarity based on the taxonomical path of ontology. For the primitive concepts, they are not completely defined in the ontology. Therefore, similarity value between primitive concepts based on the ontology structure cannot give correct degree of similarity because there have not enough inter-links connecting evaluated concepts, and similarity value will be high when ontology builder can add complete information for these concepts. Therefore, finding primitive concepts similarity based on taxonomical structure did not match with the human expert results. The evidence is shown in Table 1.

Table 1. Incomparable Similarity Values of Path-based Measure with Human Results

Concept P_1	Concept P_2	Path-based	human result
Hormonal tumor	Malignant mast cell tumor	0.2	0.6
Maternal autoimmune hemolytic anemia	Autoimmune hemolytic anemia	0.2	0.8
Atypical chest pain	Psychogenic back pain	0.3	0.6
Acute uterine inflammatory disease	Mycoplasmal pelvic inflammatory disease	0.4	0.9

After analyzing the results, path-based measure cannot give accurate values for primitive concepts because primitive concepts are defined partially in the ontology. When we compare primitive concepts, we get low similarity values because of few inter-links between them but their similarities are high in reality. Therefore, we consider primitive concepts similarity based on 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.

From this point of view, we propose a new similarity measure based on text label from the natural language processing views. We modified 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. Using context-free grammar to compute the syntactic similarity based on the noun phrase structure of concept name.

4.1. 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 [7]. We cannot omit the headword in noun phrases 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 right-most 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 give different weights for each component, the distance from the headword is considered because the nearer component to the headword should get higher weight and it has higher semantic influence on the headword than other words. As a result, the weight can be distributively estimated as shown in Table 2 and 3.

Table 2. 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 3. 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 use the Jaccard similarity [9] for two concepts similarity of headword noun phrase structure denoted by $\mathbf{sim}_{Headword}$.

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

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

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

$$= 0.43$$

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

1. Some words are lexically similar but have different meanings
   - eg: “kidney parenchyma” and “kidney beans”
   - “kidney parenchyma” is about human tissue of kidney and “kidney beans” is about one kinds of beans.
   - It cannot occur in this case 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
   - eg: 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 [6].

$$S_{synset}(P_1,P_2) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

- 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.

$$Sim(P_1, P_2) = \begin{cases} 1, & \text{if } S_{synset}(P_1, P_2) > 0 \\ 0 & \text{if } S_{synset}(P_1, P_2) = 0 \end{cases} \tag{3}$$

4.2. 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) [5]. 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, nominator is the intersection of rules and denominator is the maximum number of rules.

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

4.3. Proposed Measure

After getting similarity values from two dimensions, we consider finalize similarity values by giving different weights based on their generalization. In this case, there has a little different syntactic structures for ontology concepts. Therefore, if two concepts are exactly same syntactic structure, but different headword terms, they have so much different

meanings. But for the linguistic headword structure, it gives the accurate similarity value according to their headword position. This means that linguistic 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.

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

5. 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. From the disorder category, we evaluate 30 disease concepts using path-based measure and our proposed measure as shown in Table 4. To examine the validity of our approach, we evaluate the result of our proposed method against human expert judgement. From SNOMED CT, 30 disease concepts are selected for evaluation. Three medical doctors make a consensus on the degree of similarity of the concepts as shown in Table 4.

5.1. Discussion

Our proposed similarity measure calculates 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 approach 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 order to evaluate our proposed measure against human expert result, we compute the correlation values for our accuracy based on the results in Table 4. We got 0.83 correlation values with human expert results as shown in Table 5. Using path-based measure for primitive concepts cannot give correct similarity values because primitive concepts are not constructed fully in the ontology. So, it gets very few correlation value (0.05) with the human expert results. Moreover, we compute the error value (mean squared error) which measures the average of the errors between two results as shown in Table 6. The results show that our proposed measure gets very few error value (0.007) that very near to zero. But existing path-based measure gets larger error value (0.07) than our proposed measure. Therefore, concept name similarity is essential for primitive concept measure. 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.

Table 4. Results of degree of similarity on 30 disease concepts estimated by path-based method, our proposed method, and human expert

Concept P_1	Concept P_2	Path-based method	Proposed method	Human expert result
Hormonal tumor	Malignant mast cell tumor	0.2	0.5	0.6
Maternal autoimmune hemolytic anemia	Autoimmune hemolytic anemia	0.2	0.8	0.8
Hypertensive leg ulcer	Solitary anal ulcer	0.3	0.5	0.4
Bovine viral diarrhea	Bovine coronoviral diarrhea	0.6	0.7	0.7
Liver cell carcinoma	Blastomycosis liver	0.4	0.7	0.6
Right main coronary artery thrombosis	Superior mesenteric vein thrombosis	0.7	0.5	0.6
Acute uterine inflammatory disease	Mycoplasmal pelvic inflammatory disease	0.4	0.9	0.9
Infectious mononucleosis hepatitis	Chronic alcoholic hepatitis	0.3	0.5	0.5
Primary cutaneous blastomycosis	Primary pulmonary blastomycosis	0.7	0.7	0.6
Corneal ulcer	Acute gastrojejunal ulcer	0.5	0.4	0.6
Iodine-deficiency-related multinodular endemic goiter	Non-toxic multinodular goiter	0.8	0.8	0.8
Cerebral venous sinus thrombosis	Phlebitis cavernous sinus	1.0	0.6	0.6
Complex periorbital laceration	Third degree perineal laceration	0.3	0.5	0.5
Congenital pharyngeal polyp	Uterine cornual polyp	0.4	0.5	0.5
Mosquito-borne hemorrhagic fever	Glandular fever pharyngitis	0.4	0.5	0.5
Phakic corneal edema	Corneal epithelial edema	0.2	0.5	0.5
Coronary artery rupture	Right main coronary artery thrombosis	0.9	0.5	0.4
Rheumatic heart valve stenosis	Coronary artery stenosis	0.6	0.5	0.6
Knee pyogenic arthritis	Gonococcal arthritis dermatitis syndrome	0.9	0.4	0.4
Hereditary canine spinal muscular atrophy	Spinal cord concussion	0.5	0.3	0.5
Simple periorbital laceration	Brain stem laceration	0.5	0.4	0.5
Mite-borne hemorrhagic fever	Meningococcal cerebrospinal fever	0.4	0.6	0.5
Nasal septal hematoma	Vocal cord hematoma	0.3	0.5	0.5
Congenital cleft larynx	Congenital spastic foot	0.6	0.3	0.3
Congenital acetabular dysplasia	Short rib dysplasia	0.5	0.5	0.5
Intestinal polyposis syndrome	Ovarian vein syndrome	0.6	0.6	0.5
Extrapulmonary subpleural pulmonary sequestration	Pulmonary alveolar proteinosis	0.7	0.4	0.4
Congenital subaortic stenosis	Rheumatic aortic stenosis	0.9	0.6	0.7
Atypical chest pain	Psychogenic back pain	0.3	0.5	0.5
Puerperal pelvic cellulitis	Chronic female pelvic cellulitis	0.9	0.8	0.7

Table 5. Correlation Values between Similarity Measures and human experts

Method	Method Type	Correlation
Path-based	ontology-based measure	0.05
Proposed measure	textual annotations/ concept names	0.83

Table 6. Error Values between Similarity Measures and human experts

Method	Method Type	Error value
Path-based	ontology-based measure	0.07
Proposed measure	textual annotations/ concept names	0.007

6. Conclusion and Future Work

This paper reviews the ontology taxonomical-based similarity measure and point out the important fact for the primitive concept similarity. Primitive concepts are not well defined in the ontology so finding the primitive concepts similarity based on ontology structure gives incorrect similarity values. To overcome the weakness, this paper proposed a new primitive concept name similarity measure based on semantic and syntactic similarities and get the high correlation from human expert results.

There are some directions for our future work. Firstly, we will evaluate our approach against more number of primitive concepts. Secondly, we will evaluate our proposed measure for the defined concepts to confirm the validity of our approach on both types of concept i.e. defined concepts and primitive concepts. Finally, we will compare our approach with other ontology-based measures by analyzing benefits and usability for both defined concepts and primitive concepts.

References

[1] 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.

[2] 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.

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

[4] 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.

[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] 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.

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

[8] Suwan Tongphu and Boontawee 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.

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