

## Research Article

# An Ontology-Based Semantic Similarity Measure Considering Multi-Inheritance in Biomedicine

Fengqin Yang,<sup>1,2</sup> Yuanyuan Xing,<sup>3</sup> Hongguang Sun,<sup>1,2</sup> Tieli Sun,<sup>1,4</sup> and Siya Chen<sup>1</sup>

<sup>1</sup>School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China

<sup>2</sup>Key Laboratory of Intelligent Information Processing of Jilin Universities, Changchun 130117, China

<sup>3</sup>Students' Affairs Division, Qingdao Technological University, Qingdao 266033, China

<sup>4</sup>College of Humanities and Science, Northeast Normal University, Changchun 130117, China

Correspondence should be addressed to Hongguang Sun; sunhg889@nenu.edu.cn and Tieli Sun; suntl@nenu.edu.cn

Received 12 February 2015; Accepted 27 April 2015

Academic Editor: Chih-Cheng Hung

Copyright © 2015 Fengqin Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Computation of semantic similarity between words for text understanding is a vital issue in many applications such as word sense disambiguation, document categorization, and information retrieval. In recent years, different paradigms have been proposed to compute semantic similarity based on different ontologies and knowledge resources. In this paper, we propose a new similarity measure combining both superconcepts of the evaluated concepts and their common specificity feature. The common specificity feature considers the depth of the Least Common Subsumer (LCS) of two concepts and the depth of the ontology to obtain more semantic evidence. The multiple inheritance phenomenon in a large and complex taxonomy is taken into account by all superconcepts of the evaluated concepts. We evaluate and compare the correlation obtained by our measure with human scores against other existing measures exploiting SNOMED CT as the input ontology. The experimental evaluations show the applicability of the measure on different datasets and confirm the efficiency and simplicity of our proposed measure.

## 1. Introduction

In the last few years, the amount of available electronic information has increased sharply in many research areas such as biomedicine, education, psychology, linguistics, cognitive science, and artificial intelligence. As we know, most of the information sources are presented in unstructured or semistructured textual formats. Hence, it is an urgent issue to process the text information from a semantic perspective. Understood as the degree of taxonomical proximity, semantic similarity computes the likeness between words and plays a very important part in the above-mentioned fields such as word sense disambiguation [1], word spelling correction [2], automatic language translation [3], document categorization or clustering [4], information extraction and retrieval [5–7], detection of redundancy, and ontology learning [8, 9]. It is worth mentioning that many applications of semantic similarity computation are discussed in the biomedical domain due to the availability of numerous medical ontologies and resources that organize medical concepts into hierarchies. For

example, semantic similarity between concepts of ontologies such as Gene [10, 11] was computed with the aim of assessing protein functional similarity [6].

As mentioned above, semantic similarity is relevant to many research areas. Designing accurate computing methods is important for improving the performance of applications dependent on semantic similarity. Essentially, semantic similarity measures assess a score between a pair of words making use of the information from some predefined knowledge sources (such as ontologies or domain corpora) containing the semantic evidence. Therefore, the accuracy of semantic similarity approaches relies on the knowledge sources. So far, many semantic similarity measures have been proposed, which can be classified according to the domain knowledge itself exploited and the theoretical principles. The semantic similarity measures can be roughly divided into several categories as follows: (1) measures based on the taxonomical structure of the ontology, which are strategies estimating semantic similarity by counting the number of nodes or edges separating two concepts [12–14]; even if these methods are

the most intuitive and easy to implement, they suffer from the limitation that they work properly requiring consistent and rich ontologies; (2) measures utilizing information content (IC) of concepts, which are methods exploiting the notion of IC defined as a measure of the fruitful semantic information of concepts and computed by counting the occurrence of words in large corpora [15–17]; their shortcomings are that it is necessary to perform time-consuming analysis of corpora and that the IC values depend on the considered corpora; (3) measures using the amount of cooccurrences between word contexts, which are approaches constructing context vectors of concepts by extracting contextual words (within a fixed window of context) from a corpus of textual documents including the evaluated concepts and computing the similarity of concepts as the cosine of the angle between their context vectors [11, 18, 19]. Similar to the methods motioned in category (2), the availability and suitability of corpora affect the applicability of these measures.

Usually, these measures can obtain good performance when we employ large and general purpose knowledge bases like WordNet [20]. Some of them have been applied to the biomedical field using domain information extracted from clinical data or relevant medical ontologies such as SNOMED CT (<https://uts.nlm.nih.gov/home.html>) [21, 22] or MeSH (<http://www.nlm.nih.gov/mesh/meshhome.html>) [22, 23] in the Unified Medical Language System (UMLS) [22] and authors compared these measures and analyzed and evaluated them over certain datasets to determine their advantages and limitations with respect to the background knowledge source [24–26].

In this paper, firstly, we review and investigate different measures for semantic similarity computation. Then, we propose a new measure considering the multiple inheritance in ontologies and the common specificity feature of the evaluated concepts in order to obtain a more accurate similarity between concepts. Finally, we evaluate the proposed measure using two datasets of biomedical term pairs scored for similarity by human experts and exploiting SNOMED CT as the input ontology. We compare the correlation obtained by our measure with human scores against other measures. The experimental evaluations confirm the efficiency of the proposed measure.

Besides Section 1, the paper is organized as follows. Section 2 investigates the basic methods for semantic similarity including the taxonomy-based measures, the IC-based measures, and the context vector measures. Section 3 presents the proposed measure for semantic similarity and its main advantages. Section 4 evaluates and compares the measure against the analyzed measures using SNOMED CT as the input ontology. Section 5 analyzes and discusses the experimental results. The final section is the conclusions of the paper.

## 2. Existing Measures for Computing Semantic Similarity

The existing measures for semantic similarity are discussed as follows.

**2.1. Measures Based on the Taxonomical Structure.** The simplest way of computing similarity for concepts is the measure based on path length developed by Rada et al. [13]. The measure quantifies the shortest distance between the two concept nodes  $c_1$  and  $c_2$ :

$$\text{dis}(c_1, c_2) = N_1 + N_2, \quad (1)$$

where  $N_1$  and  $N_2$  stand for the minimum number of is-a links from  $c_1$  and  $c_2$  to their LCS, respectively.

Wu and Palmer [14] introduced a measure based on path length that considers the depth of the concepts only in the hierarchy. It is based on the assumption that concepts lower down in the taxonomy are more similar than those higher up:

$$\text{sim}_{\text{W\&P}}(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}, \quad (2)$$

where  $N_3$  is the number of is-a relations from the LCS of the evaluated concepts to the root of the ontology. And the similarity value ranges from 1 (for identical concepts) to 0.

Leacock and Chodorow [12] proposed a measure for similarity in which the shortest path length between two concepts was scaled by twice the maximum depth  $D$  of the hierarchy:

$$\text{sim}_{\text{L\&C}}(c_1, c_2) = -\log\left(\frac{N_1 + N_2 + 1}{2 \times D}\right). \quad (3)$$

Besides, there are other measures for semantic similarity based on structure. Li et al. [27] developed a measure combining the depth of the ontology and the shortest path:

$$\text{sim}_{\text{Li}}(c_1, c_2) = e^{-\alpha \text{path}(c_1, c_2)} \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \quad (4)$$

where  $h$  is the minimum depth of the LCS in the hierarchy,  $\text{path}(c_1, c_2)$  is the shortest path between two concepts, and  $\alpha \geq 0$  and  $\beta > 0$  stand for the contribution of the shortest path and the depth, respectively. The optimal parameters for the measure were  $\alpha = 0.2$  and  $\beta = 0.6$ .

Al-Mubaid and Nguyen [24] proposed a cluster-based measure combining path length and common specificity that considers the depth of the LCS of two concepts and the depth  $D$  of ontology. They defined the clusters as the branches of the ontology with respect to the root node. The common specificity of concepts  $c_1$  and  $c_2$  is defined as follows:

$$\text{CSpec}(c_1, c_2) = D_c - \text{Depth}(\text{LCS}(c_1, c_2)), \quad (5)$$

where  $D_c$  is the depth of the cluster including concepts  $c_1$  and  $c_2$ . Thus the  $\text{CSpec}(c_1, c_2)$  feature determines the ‘‘common specificity’’ of two concepts in the cluster. The smaller the common specificity value of two concept nodes is, the more information they share. Thus, they will be more similar. The semantic distance measure is defined as follows:

$$\text{SemDist}_{\text{A\&N}}(c_1, c_2) = \log\left((\text{Path} - 1)^\alpha \times (\text{CSpec})^\beta + k\right), \quad (6)$$

where  $\alpha$  and  $\beta$  ( $\alpha > 0, \beta > 0$ ) are contribution factors of two features,  $k$  is a constant which must be greater than or equal to 1 to insure that the distance is positive and the combination is nonlinear, and Path is the length of the shortest path between the two concept nodes.

Batet et al. [25] proposed a similarity measure which takes into account all the superconcepts (subsumers of the evaluated terms) belonging to all the possible paths between the concept nodes and defined the measure as the ratio between the amount of nonshared information and the sum of shared and nonshared information. The similarity measure of two concepts considering multiple inheritance is defined as follows:

$$\begin{aligned} \text{sim}_{\text{B\&S}}(c_1, c_2) \\ = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|}, \end{aligned} \quad (7)$$

where  $T(c_i) = \{c_j \in C \mid c_j \text{ is superconcept of } c_i\} \cup \{c_i\}$  and  $C$  is the concept set.

From the measures introduced above, we can get a conclusion that the LCS of two concepts plays a vital part in the computation for semantic similarity of concepts. For some measures, it is enough that just only ontology is used as background knowledge (no corpus with domain data is needed), which makes them heavily depend on the ontology itself. Besides, the minimum path (shortest path) also plays an important part. For large ontologies, such as SNOMED CT, there is a phenomenon that one or both concepts inherit from several is-a hierarchies. In this case, Al-Mubaid and Nguyen [24] used the minimum path to get the maximum semantic similarity. However, it may omit much other available taxonomical knowledge. Here, multiple possible paths can exist between any two concepts, but only the shortest one is selected among those paths. To solve the problem, we can take all the superconcepts into account and try to get more semantic evidence in the case of multiple inheritance, which makes the measure for semantic similarity more accurate.

In a view of an independent domain, in order to get high accuracy, most path-based measures rely on large and general purpose taxonomy. Usually, researchers may choose WordNet to apply these measures because of its perfect structure. However, the coverage of biomedical terms in WordNet is so limited that the accuracy of similarity assessments for medical terms is poor [11, 28]. So Pedersen et al. [11], Al-Mubaid and Nguyen [24], and Batet et al. [25] adopted these measures to the biomedical domain by exploiting SNOMED CT as the input ontology.

**2.2. Measures Based on Information Content.** These measures evaluate the similarity of concepts depending on the amount of shared information between two concepts. According to the information theory, concepts are evaluated by their IC. IC can be considered as a measure which quantifies the amount of information that a given concept expresses when appearing in taxonomy. Resnik [17] stated that concept semantic similarity depends on the amount of shared information between them.

In Resnik's seminal work [17], IC is computed according to  $p(c)$  representing the probability of occurrence of a concept  $c$  in a corpus:

$$\text{IC}(c) = -\log p(c). \quad (8)$$

Usually, in a general context,  $p(c)$  estimation is severely hampered due to textual ambiguity and data sparseness problems [29]. In fact, the tagged corpora about domain information like biomedicine are limited. Authors [15, 17] estimated concept appearance from SemCor [30] that is a semantically tagged text consisting of 100 passages from the Brown Corpus based on WordNet. Because the manual tagging scheme is based on the fine grained structure of word senses covered by WordNet, the  $p(c)$  estimation is accurate but is limited to the coverage of corpora that covered less than 13% of the available word senses in WordNet.

To guarantee the consistency of similarity computation, coherence of  $p(c)$  computation based on taxonomical structure should be taken into account. Meanwhile, to compute the value, both the explicit appearances of concept  $c$  and its specializations must be considered. Thus, Resnik [17] proposed the measure for calculating  $p(c)$  showed as formula (9). Consider

$$p(c) = \frac{\sum_{n \in w(c)} \text{count}(n)}{N}, \quad (9)$$

where  $w(c)$  is the set of words subsumed by concept  $c$  and  $N$  is the total number of observed corpus terms excluding those that are not subsumed by any WordNet class.

Resnik [17] mentioned concepts in a lower level which are usually more specialized and share more information represented by LCS of both concepts in taxonomy. So the more the IC of the subsumer of concepts is, the more similar the concepts are. Based on the premise above, Resnik measures the similarity as the IC of the LCS of concepts:

$$\text{sim}_{\text{res}}(c_1, c_2) = \text{IC}(\text{LCS}(c_1, c_2)). \quad (10)$$

To tackle the problem of Resnik's measure that the similarity value will be the same if any two concepts have the same LCS, both studies by Jiang and Conrath [15] and Lin [16] improved Resnik's measure.

Lin measured the similarity as the ratio between the IC of the LCS of the two concepts and the summation of the IC of the two concepts:

$$\text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \times \text{IC}(\text{LCS}(c_1, c_2))}{(\text{IC}(c_1) + \text{IC}(c_2))}. \quad (11)$$

Jiang and Conrath calculated the dissimilarity between concepts illustrating the similarity of concepts with formula (12). Consider

$$\begin{aligned} \text{dis}_{\text{J\&C}}(c_1, c_2) &= (\text{IC}(c_1) + \text{IC}(c_2)) - 2 \\ &\times \text{IC}(\text{LCS}(c_1, c_2)). \end{aligned} \quad (12)$$

As mentioned above, the IC-based similarity assessments will not get the best performance when we use a general purpose corpus such as WordNet or SemCor due to their limited

coverage of biomedical terms. For this reason, Pedersen et al. [11] applied these measures to the biomedical domain by exploiting SNOMED CT and computing the IC of concepts using Mayo Clinic Corpus of Clinical Notes as a domain corpus.

**2.3. Context Vector Measures Computing Semantic Relatedness.** Patwardhan and Pedersen [19] proposed a measure of semantic relatedness that represents a concept with a context vector which performs more flexible than similarity measurements, since the information source for the context vectors is a raw corpus of text and the concepts do not need to be connected via a path of relations in ontology. They built gloss vectors corresponding to each concept in WordNet using the cooccurrence information along with the WordNet definitions. In their experiments, the glosses seemed to contain content rich terms. They would distinguish various concepts much better than text drawn from more generic corpus if authors chose the WordNet glosses. And the WordNet glosses can be viewed as a corpus of contexts consisting of about 1.4 million words. The gloss vector measure got the highest correlation with respect to human judgment using different benchmarks [19].

Pedersen et al. [11] constructed cooccurrence vectors that represent the contextual profile of concepts and applied the measure to the biomedical field. In their study, they created context vectors corresponding to each concept in SNOMED CT using a set of word vectors. The word vectors for all words occurring in the clinical notes referred to here were produced with the Mayo Clinic Corpus of Clinical Notes. In these vectors, the window size of the context is one line of the text.

Then, the semantic relatedness of two concepts  $c_1$  and  $c_2$  is computed as the cosine of the angle between their context vectors with formula (13). Consider

$$\text{rel}_{\text{vector}}(c_1, c_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|}, \quad (13)$$

where  $v_1$  and  $v_2$  are the context vectors corresponding to  $c_1$  and  $c_2$ , respectively. Note that the context vector measure has different performance according to the different choices about clinical notes in the experiments. In other words, this measure depends heavily on the availability and quality of the corpora.

### 3. Proposed Measure for Computing the Semantic Similarity

According to the analysis of similarity measures above, we know measures based on IC or context vectors depend on corpora. As a matter of fact, corpora consist of unstructured or semistructured textual data. Corpora need to be preprocessed to obtain enough semantic information which brings a heavy computational burden. In the biomedical domain, it is very difficult to get enough clinical data due to the sensitivity of patient data, which may cause data sparseness problems. For these reasons, the applicability of these measures may be hampered by the availability of enough suitable data. On

the contrary, path-based measures only use the structure of ontologies and do not require preprocessing of text data, which makes them have low computational complexity. But every coin has two sides. Path-based measures are very simple and it cannot get more semantic evidence to perform better than other measures like IC-based measures and context vector relatedness measures.

From the path-based measures, we know that when the path length between each of concepts and their LCS is calculated, only the minimum path length is found out and kept for use, even if all paths may be calculated. However, in the taxonomy, one or both concepts may inherit from several is-a hierarchies. For this reason, there exists a case of multiple inheritance [25, 31], especially in large and complex taxonomies (e.g., SNOMED CT) including thousands of interrelated concepts in the hierarchies. For example, in Figure 1, if we choose the minimum distance among all the paths between  $c_1$  and  $c_2$ , the contributions of the noncommon superconcepts of the two concepts in the taxonomy are omitted, which affects the accuracy of the measures.

In this paper, we improve significantly the measure of Al-Mubaid and Nguyen [24] and propose a new modified measure for semantic similarity combining both superconcepts of the evaluated concepts and common specificity feature which can capture more semantic evidence. The proposed measure can achieve better performance than other measures based on structure and keep their simplicity.

To take into account contributions of all noncommon superconcepts of the evaluated concepts, we consider the concepts themselves and all their noncommon superconcepts instead of the minimum path length to capture more semantic evidence for similarity. Besides, in our measure, we also consider the common specificity feature of the concepts nodes scaled by the depth of the concept nodes and the depth of their LCS. Thus, we combine noncommon superconcepts with common specificity in order to get more semantic information for computing similarity between two concepts.

Let  $c_i$  stand for  $i$ th concept of an ontology. Then  $N(c_i)$  is defined as the set of all superconcepts of  $c_i$  including  $c_i$  itself. Thus the number of noncommon superconcepts of concepts can be defined as follows:

$$\text{NonComSub}(c_1, c_2) = |N(c_1) \cup N(c_2)| - |N(c_1) \cap N(c_2)|. \quad (14)$$

Here the NonComSub value can be an indication of the path length of the two concepts. For example, the number of noncommon superconcepts for concepts  $c_1$  and  $c_2$  is 3 in Figure 1.

On the other hand, the common specificity feature is defined as follows [24]:

$$\text{ComSpec}(c_1, c_2) = D - \text{Depth}(\text{LCS}(c_1, c_2)), \quad (15)$$

where  $D$  is the depth of the ontology and the ComSpec feature determines the common specificity of two evaluated concepts. The smaller the ComSpec value of two concepts, the more information they share, and thus the more similar they are.

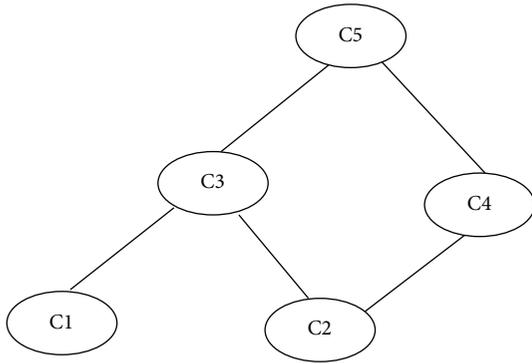


FIGURE 1: Ontology with multi-inheritance.

Then, we use logarithm function of NonComSub and ComSpec to represent semantic distance which is inverse to semantic similarity. Therefore, the semantic distance between concepts  $c_1$  and  $c_2$  is defined as follows:

$$\text{SemD}(c_1, c_2) = \log(\text{NonComSub} \times \text{ComSpec} + 1). \quad (16)$$

It is worth mentioning that any concept can be compared with itself. In this case, the semantic distance is 0. This measure can be applied to all concepts and does not need to check whether or not the two concepts compared are distinct.

#### 4. Evaluation

Measures of semantic similarity are usually evaluated by comparing the computed similarity values of the measures against the human judgments using correlation coefficient. The higher the correlation value against the human experts' similarity scores are, the better the measure is.

In the biomedical field, there are no standard human rating datasets for semantic similarity like manually rated concept sets created by Rubenstein and Goodenough [32] and Miller and Charles [33]. Pedersen et al. [11] stated that it is necessary to choose sets of words manually scored for the evaluation of concept semantic similarity measures in biomedicine. In their research, they created a set of 30 concept pairs regarding medical disorders with the help of Mayo Clinic experts. The set was annotated by three physicians and nine medical coders. All three physicians are specialists in the area of rheumatology. Finally, after a series of processing, the average of the similarity values of 30 concept pairs was normalized in a scale between 1 and 4. The average correlation between physicians is 0.68 while the average correlation between medical coders is 0.78. The 30 medical term pairs with averaged experts' similarity scores are showed in Table 1.

Pedersen et al. [11] applied the standard to evaluate the path-based and the IC-based measures by exploiting SNOMED CT taxonomy as domain ontology and the Mayo Clinical Corpus and Thesaurus as corpora, respectively. Medical coders had a better understanding about the notion of similarity because they were pretrained during the construction of the original dataset. So medical coders' ratings seem to reproduce better the concept of (taxonomic) similarity,

whereas physicians' ratings seem to represent a more general concept of (taxonomic and nontaxonomic) relatedness. Al-Mubaid and Nguyen [24] just compared their concept similarity value against coders' ratings, while Batet et al. [25] made a comparison between concept similarity value obtained and scores of both of them.

In this paper, in order to compare the performance of our measure with other measures in the biomedical domain objectively, we used the set of 30 concept pairs from literature [11] (called Dataset 1) and the set of 36 biomedical term pairs from literature [6] (called Dataset 2) as experimental datasets. The 36 medical term pairs in Dataset 2 with averaged human' similarity scores are showed in Table 2, in which the human scores are the average evaluated scores of reliable doctors. We use UMLS Knowledge Source (UMLS) browser (<https://uts.nlm.nih.gov/home.html>) for SNOMED CT to get information on the terms in the two datasets. The comparative results using Dataset 1 with respect to both physicians and medical coders and results using Dataset 2 with respect to human are shown in Table 3. Note that our measures compute semantic distance while human ratings represent similarity. So a linear transformation should be performed.

#### 5. Discussion

For Dataset 1, there are 29 out of 30 concept pairs in SNOMED CT and the average correlation between physicians is 0.68 while the average correlation between medical coders is 0.78. The evaluation results show that the path-based similarity measures obtain lower correlations than 0.36 and 0.66 for physicians and coders, respectively. It indicates that the accuracy of path-based measures is limited. These measures utilize the minimum path length without considering multiple inheritances, which causes much useful semantic evidence to be ignored. For the IC-based measures, they improve the results of most measures based on path length generally with the highest value 0.75 for coders and 0.60 for physicians. And the lowest correlation is 0.62 for coders, which outperforms most structure-based measure with exception of SemDist<sub>A&N</sub> (0.66) and Sim<sub>B&S</sub> (0.79). The coverage of biomedical terms in domain corpora is limited. Due to the high dependency on the availability of domain data, the accuracy of the IC-based similarity approaches is hampered.

The measure Sim<sub>B&S</sub> shows good performance considering both common and uncommon information of the evaluated concepts. However, when using Sim<sub>B&S</sub> to compute semantic similarity, we should check if the two concepts are the same; otherwise we would get an infinitely large similarity value.

With respect to the context vector measure, there are four cases with changing corpus size and corpus selection. From the results, we can see that the best correlations are 0.84 for the physicians and 0.75 for the coders under the condition of 1 million notes involving only the diagnostic section. That is, the correlation value with the context vector measure is higher than ours only if 1 million notes are used to create the vectors. Moreover, the data corpus used to create vectors was constructed by physicians of the same

TABLE 1: Dataset 1: 30 medical term pairs with physicians' and coders' similarity scores [11].

Term 1	Term 2	Physician ratings (Averaged)	Coder ratings (Averaged)
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpal tunnel syndrome	Osteoarthritis	2.0	1.1
Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulitis	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlipidemia	Metastasis	1.0	1.0

clinic and so there were some limitations in the way of interpreting and formalizing knowledge. Thus the context vector measure strongly depends on the amount and quality of the background corpus. The measure can show good performance just in the particular situation and domain. If we want to obtain higher correlation, we should try best to make preparations in choosing the size and quality of the information sources. The results show that the accuracy of the approach decreases for other corpus configurations noticeably. For example, when 100,000 notes involving all sections are used, the correlations drop to 0.41 for physicians and 0.53 for coders.

The correlations obtained by our measure are 0.67 for physicians and 0.77 for coders while the average correlation between physicians is 0.68 and the average correlation between medical coders is 0.78. The proposed measure has made obvious improvement to  $\text{SemDist}_{A\&N}$  (0.77 versus 0.66). The correlation for physicians is higher than other measures except for context vector measure and the correlation for coders is higher than other measures except for

$\text{Sim}_{B\&S}$  (0.79). And the correlation is 0.75 considering both sets of experts, which is rather high among all the measures mentioned in this paper. Our measure gets higher correlation values than all the IC-based measures shown in Table 3.

From Table 3, we can see that the correlation values for coders are always higher than that for physicians except for context vector measure which explains that medical coders' ratings with more pretraining are more reliable than physicians' ratings. As a result, many similarity measures compare similarity against similarity scores of medical coders to obtain much better correlations.

For Dataset 2, we can find 34 out of 36 concept pairs in SNOMED CT. We compare our measure with other structure-based similarity measures with respect to human scores. The correlation of Al-Mubaid and Nguyen measure is 0.735 which is higher than other measures. However, the correlation value obtained by our measure is 0.774. The comparative result (0.774 versus 0.735) shows that the proposed measure outperforms other measures shown in Table 3.

TABLE 2: Dataset 2: 36 medical term pairs with averaged human similarity scores [6].

Term 1	Term 2	Human
Anemia	Appendicitis	0.031
Dementia	Atopic dermatitis	0.062
Bacterial pneumonia	Malaria	0.156
Osteoporosis	Patent ductus arteriosus	0.156
Amino acid sequence	Antibacterial agents	0.156
Acquired immunodeficiency syndrome	Congenital heart defects	0.062
Otitis media	Infantile colic	0.156
Meningitis	Tricuspid atresia	0.031
Sinusitis	Mental retardation	0.031
Hypertension	Kidney failure	0.500
Hyperlipidemia	Hyperkalemia	0.156
Hypothyroidism	Hyperthyroidism	0.406
Sarcoidosis	Tuberculosis	0.406
Vaccines	Immunity	0.593
Asthma	Pneumonia	0.375
Diabetic nephropathy	Diabetes mellitus	0.500
Lactose intolerance	Irritable bowel syndrome	0.468
Urinary tract infection	Pyelonephritis	0.656
Neonatal jaundice	Sepsis	0.187
Sickle cell anemia	Iron deficiency anemia	0.437
Psychology	Cognitive science	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial ischemia	Myocardial infarction	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary valve stenosis	Aortic valve stenosis	0.531
Failure to thrive	Malnutrition	0.625
Breast feeding	Lactation	0.843
Antibiotics	Antibacterial agents	0.937
Seizures	Convulsions	0.843
Pain	Ache	0.875
Malnutrition	Nutritional deficiency	0.875
Measles	Rubeola	0.906
Chicken pox	Varicella	0.968
Down syndrome	Trisomy 21	0.875

Through the observation of the experimental results on two datasets, we find that our measure performs much better than almost all the similarity measures including the

path-based measures and the IC-based measures. We make significant improvements to the measure of Al-Mubaid and Nguyen in the case of multi-inheritance which exists in SNOMED CT.

In addition, we adopt the Pearson correlation coefficient ( $p$  value) as a measure of the strength of the relation between human ratings of similarity and computational values [30]. The smaller the  $p$  value is, the more significant the relation is. In all cases, the  $p$  values for our results on both Dataset 1 and Dataset 2 are less than 0.001 (0.1% chance), which explains that the correlation values using our measure are significant statistically.

By all accounts, our measure is only based on an ontology structure and it can provide a comparatively high accuracy without any dependency on data preprocessing and availability. Meanwhile, it keeps the simplicity of structure-based measures and overcomes their shortcomings that multi-inheritance phenomenon does not fully considered. The measure nonlinearly combines the evaluated concepts' non-common information considering the different taxonomical hierarchies and their common specificity feature, which has made obvious improvement to other measures. Note that there exists the case of multiple inheritance in which concepts may be subsumed by several superconcepts in the largest and most widely used knowledge source such as SNOMED CT, MeSH in the UMLS, or WordNet. From the experiments, our measure has very good performance in ontologies with multi-inheritance. When the input ontology does not have the multiple inheritance of concepts, the measure can also be accurate. Therefore, the accuracy and application of our approach are very impressive and significant especially in the biomedical domain.

## 6. Conclusions

In the paper, we propose a similarity measure that nonlinearly combines the evaluated concepts' noncommon information considering the different taxonomical hierarchies with their common specificity feature. The measure keeps the simplicity without parameter tuning. In the experiments, we use SNOMED CT, a large and detailed ontology with multiple inheritance between concepts, as input ontology. The experimental results show that our measure has rather high correlation values with respect to both physicians and coders and the measure outperforms most approaches based on taxonomical structure and IC and context vectors.

As we know, recently, measuring semantic similarity of concepts within multiple ontologies has become more and more important. As future work, we will extend the measure to multiple ontologies such as SNOMED CT, MeSH, and WordNet.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

TABLE 3: Correlations across measures for physicians, coders, and both on 29 pairs of Dataset 1 and correlations across measures for human on 34 pairs of Dataset 2 using SNOMED CT as ontology.

Measure	Dataset 1			Dataset 2
	Physicians	Coders	Both	Human
Path length	0.36	0.51	0.48	0.586
Leacock and Chodorow	0.35	0.50	0.47	0.677
Wu and Palmer	N/A	0.29	N/A	0.686
Li et al.	N/A	0.37	N/A	0.694
Choi and Kim	N/A	0.15	N/A	0.440
Al-Mubaid and Nguyen (SemDist <sub>A&amp;N</sub> )	N/A	0.66	N/A	0.735
Montserrat Batet et al. (Sim <sub>B&amp;S</sub> )	0.60	0.79	0.73	N/A
Resnik	0.45	0.62	0.55	N/A
Lin	0.60	0.75	0.69	N/A
Jiang and Conrath	0.45	0.62	0.55	N/A
Context vector (1 m notes, diagnostic section)	0.84	0.75	0.76	N/A
Context vector (1 m notes, all sections)	0.62	0.68	0.69	N/A
Context vector (100,000 notes, diagnostic section)	0.56	0.59	0.60	N/A
Context vector (100,000 notes, all sections)	0.41	0.53	0.51	N/A
Our measure	0.67	0.77	0.75	0.774

## Acknowledgments

This paper was sponsored by Jilin Provincial Science and Technology Department of China (Grant nos. 20130206041GX and 20120302), Jilin Province Development and Reform Committee of China (Grant no. 2013C036-5, [2013]779), Changchun Science and Technology Bureau of China (Grant no. 14KT009), and the Doctoral Program of Higher Education of China (no. 20110043110011), respectively. The authors appreciate the help of Professor David Sánchez in their research. He provided them with useful information and friendly advice.

## References

- [1] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation," in *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, February 2003*, pp. 241–257, Springer, Berlin, Germany, 2003.
- [2] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.
- [3] R. L. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.
- [4] S. Aseervatham and Y. Bennani, "Semi-structured document categorization with a semantic kernel," *Pattern Recognition*, vol. 42, no. 9, pp. 2067–2076, 2009.
- [5] J. Atkinson, A. Ferreira, and E. Aravena, "Discovering implicit intention-level knowledge from natural-language texts," *Knowledge-Based Systems*, vol. 22, no. 7, pp. 502–508, 2009.
- [6] A. Hliaoutakis, "Semantic similarity measures in the MESH ontology and their application to information retrieval on Medline," Tech. Rep., Technical University of Crete (TUC), Department of Electronic & Computer Engineering, 2005.
- [7] M. Stevenson and M. A. Greenwood, "A semantic approach to IE pattern induction," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, pp. 379–386, Ann Arbor, Mich, USA, June 2005.
- [8] D. Sánchez and A. Moreno, "Learning non-taxonomic relationships from web documents for domain ontology construction," *Data & Knowledge Engineering*, vol. 64, no. 3, pp. 600–623, 2008.
- [9] D. Sánchez, *Domain ontology learning from the web [Ph.D. thesis]*, VDM Publishing, Saarbrücken, Germany, 2008.
- [10] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between gene ontology terms," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 137–152, 2007.
- [11] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics*, vol. 40, no. 3, pp. 288–299, 2007.
- [12] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed., pp. 265–283, MIT Press, Cambridge, Mass, USA, 1998.
- [13] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.
- [14] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pp. 133–138, ACM, Las Cruces, NM, USA, 1994.
- [15] J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of the International Conference on Research in Computational Linguistics*, pp. 19–33, Taipei, Taiwan, 1997.
- [16] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, pp. 296–304, Morgan Kaufmann, Madison, Wis, USA, 1998.
- [17] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International*

- Joint Conference on Artificial Intelligence*, pp. 448–453, Montreal, Canada, 1995.
- [18] S. Patwardhan, *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness [M.S. Science thesis]*, Department of Computer Science, University of Minnesota, Duluth, Minn, USA, 2003.
- [19] S. Patwardhan and T. Pedersen, “Using WordNet-based context vectors to estimate the semantic relatedness of concepts,” in *Proceedings of the EACL Workshop*, pp. 1–8, Trento, Italy, 2006.
- [20] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Mass, USA, 1998.
- [21] K. A. Spackman, “SNOMED CT milestones: endorsements are added to already-impressive standards credentials,” *Healthcare Informatics*, vol. 21, no. 9, pp. 54–56, 2004.
- [22] R. Kleinsorge, C. Tilley, and J. Willis, *Unified Medical Language System (UMLS) Basics*, 2000, [http://www.nlm.nih.gov/research/umls/pdf/UMLS\\_Basics.pdf](http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf).
- [23] S. J. Nelson, D. Johnston, and B. L. Humphreys, “Relationships in medical subject headings,” in *Relationships in the Organization of Knowledge*, pp. 171–184, K. A. Publishers, 2001.
- [24] H. Al-Mubaid and H. A. Nguyen, “A cluster-based approach for semantic similarity in the biomedical domain,” in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '06)*, pp. 2713–2717, IEEE, New York, NY, USA, August–September 2006.
- [25] M. Batet, D. Sánchez, and A. Valls, “An ontology-based measure to compute semantic similarity in biomedicine,” *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 118–125, 2011.
- [26] G. Pirró, “A semantic similarity metric combining features and intrinsic information content,” *Data and Knowledge Engineering*, vol. 68, no. 11, pp. 1289–1308, 2009.
- [27] Y. Li, Z. A. Bandar, and D. McLean, “An approach for measuring semantic similarity between words using multiple information sources,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.
- [28] A. Burgun and O. Bodenreider, “Comparing terms, concepts and semantic classes in WordNet and the unified medical language system,” in *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 77–82, Pittsburgh, Pa, USA, June 2001.
- [29] E. Brill, “Processing natural language without natural language processing,” in *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 360–369, Mexico City, Mexico, 2003.
- [30] G. Miller, C. Leacock, R. Teng, and R. T. Bunker, “A semantic concordance,” in *Proceedings of the ARPA Workshop on Human Language Technology Association for Computational Linguistics*, pp. 303–308, Princeton, NJ, USA, March 1993.
- [31] D. Sánchez, M. Batet, and D. Isern, “Ontology-based information content computation,” *Knowledge-Based Systems*, vol. 24, no. 2, pp. 297–303, 2011.
- [32] H. Rubenstein and J. B. Goodenough, “Contextual correlates of synonymy,” *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.
- [33] G. A. Miller and W. G. Charles, “Contextual correlates of semantic similarity,” *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.



# Hindawi

Submit your manuscripts at  
<http://www.hindawi.com>

