An Intrinsic Information Content Based Semantic Similarity Measure Considering The Disjoint Common Subsumers Of Concepts Of An Ontology

Abhijit Adhikari[a,*], Biswanath Dutta[b], Animesh Dutta[c], Deepjyoti Mondal[d], and Shivang Singh[e]

[a]Department of CSE, NIT Durgapur, West Bengal, India 713209, E-mail: abhijitbitmesra@gmail.com
[*]Corresponding Author
[b]DRTC, Indian Statistical Institute, Bangalore, Karnataka, India 560059, E-mail: dutta2005@gmail.com
[c]Department of CSE, NIT Durgapur, Durgapur, West Bengal, India 713209, E-mail: animeshnit@gmail.com
[d]Department of CSE, NIT Durgapur, Durgapur, West Bengal, India 713209, E-mail: djmdeveloper060796@gmail.com
[e]Amdocs, India, E-mail: shivangsingh777@gmail.com

Abstract

Finding similarity between concepts based on semantics has become a new trend in many applications (e.g. biomedical informatics, natural language processing). Measuring the Semantic Similarity with higher accuracy is a challenging task. In this context, Information Content (IC) based Semantic Similarity (SS) measure has gained popularity over the others. The notion of IC evolves from the science of information theory. Information theory has very high potential to characterize the semantics of concepts. Designing an IC based SS framework comprises (a) an IC calculator, and (b) a SS calculator. In this paper, we propose a generic intrinsic IC based semantic similarity calculator. We also introduce here a new structural aspect of an ontology called DCS (Disjoint Common Subsumers) that plays a significant role in deciding the similarity between two concepts. We evaluate our proposed similarity calculator with the existing intrinsic IC based similarity calculators as well as corpora dependent similarity calculators using several benchmark datasets. The experimental results show that the proposed similarity calculator produces a high correlation with human evaluation over the existing state of the art IC based similarity calculators.

Keywords: Semantic similarity; Information theory; Knowledge based measure; Intrinsic information content based measure; Ontology; Semantic similarity benchmark based on WordNet, SNOMED-CT, MeSH.

Introduction

Finding similarity between concepts is a long-standing research issue for many years both in artificial intelligence and cognitive science. It has a great significance in various applications like word sense disambiguation, information extraction, ontology merging etc. (Adhikari, Singh, Mondal, Dutta, & Dutta, 2016).

In the literature, we find several similarity measurement techniques. They are broadly classified into three (Harispe, Ranwez, Janaqi, & Montmain, 2013): distributional measure, knowledge based measure and hybrid approach. Distributional measure relies on the corpus and is mostly applied for measuring the relatedness between concepts. The primary limitation of this approach is the words to be compared must appear in the corpus at least few times (Harispe et al., 2013). Knowledge based approach relies on the user defined resources like taxonomies, thesauri, ontology or encyclopedias. Due to the availability of several knowledge sources in different domains, knowledge based approach gained popularity over the distributional approach. There are mainly three kinds of knowledge based approaches observed and these are edge based (Aouicha & Taieb, 2015), feature based (Harispe et al., 2013), and information theory based (i.e. IC based) (Harispe et al., 2013). Among these, information theory based approach is the most popular one (Sánchez & Batet, 2012; Seco, Veale, & Hayes, 2004; Sánchez, Batet, & Isern, 2011; Zhou, Wang, & Gu, 2008; Meng, Gu, & Zhou, 2012). Hybrid approach combines both the distributional and knowledge-based strategies (Harispe et al., 2013).

Existing researches (Adhikari, Singh, Dutta, & Dutta, 2015; Sánchez & Batet, 2012) show that semantic similarity (SS) measures based on IC give better accuracy than the non-IC based measures. Earlier IC based SS measures rely totally on the corpus. In those measures, how many number of times a concept or any of its instance appears in a corpus has become the base for calculating IC of a concept. Such techniques are called corpus based IC calculation measures. But those measures have data sparsity problem. Apart from this, to design a tagged corpora, an immense amount of manual efforts are needed. To overcome such issues, a completely ontology based approach has been evolved called intrinsic IC calculation approach. Existing researches reveal that corpora based IC dependent SS measures provide significantly lower correlation values with human evaluation than its intrinsic counterpart (Sánchez & Batet, 2012; Sánchez et al., 2011). The current work focuses on designing an intrinsic IC based SS calculator (model).

Information theory based SS finding is a two folded process (Adhikari et al., 2015). The first step is to calculate Information Content (IC) of each concept. In the second step, based on those IC values, the similarity between the concepts is calculated. The existing intrinsic IC based information theoretic similarity calculation models primarily focus on the various aspects of the underlying ontology for measuring the similarity, for instance, the distance between two concepts (Jiang & Conrath, 1997), features of the concepts (Pirró & Euzenat, 2010) and so forth. The core of these existing models is the Least Common Subsumer (LCS). It plays a significant role in measuring the similarity between the concepts (Resnik, 1995; Jiang & Conrath, 1997; Lin, 1998; Sánchez & Batet, 2011, 2011; Pirró, 2009). Nonetheless, LCS based similarity models have some limitations. In LCS based models, some of the common ancestors which contribute different conceptual dimensions to the underlying concepts c_i, c_j (where, c_i and c_j are any two concepts under an ontology) are completely ignored (discussed in details in “Proposed Approach”). Thus the underlying concepts miss the bigger scope to be judged for their similarity. As a solution, we introduce here a brand new aspect called DCS (Disjoint Common Subsumers). It covers nearing all the structural aspects of an underlying ontology. It is evident from our experimental result (see “Experiments”section) that the similarity models based on DCS produce a better result compared to the LCS based similarity models.

The main contributions of this work are:

- We propose an intrinsic IC based information theoretic SS calculator for measuring the SS between two concepts of an ontology. To build our similarity model, we use our proposed Disjoint Common Subsumers (DCS). We also propose an algorithm for finding the members of DCS. The accuracy of the proposed SS calculator is assessed thoroughly using a variety of benchmark datasets.
- We also bring three different ontologies, namely, WordNet [1], SNOMED-CT [2] and MeSH (Hliaoutakis, 2005) from different domains under one umbrella to evaluate our proposed SS calculation model as well as state of the art models.

The rest of the paper is organized as follows: in the next section, we describe some notations and definitions that are used throughout this paper. Next, we discuss some of the previous works in the related domain. Following this, we discuss some limitations of LCS based SS model. Then, we illustrate the core element of our proposed similarity model i.e., Disjoint Common Subsumers, followed by an algorithm for finding DCS. Next, we discuss our proposed SS calculator. Then we describe our experiments which include the task description, experimental setup, and analysis of results followed by a discussion. We conclude the paper with some future scopes.

Notations and Definitions

O represents an ontology which is a connected graph $G(V, E)$. V represents set of concepts (c) and E represents relation R between concepts. It has a special node (i.e. a concept) designated as “root”. $IC(c) \mid c$ is a concept, denotes IC of any concept and $IC(c)$ ranges between 0 and 1. $IC(root)$ in any ontology is considered “0” as it is the most abstract conceptualization among all other nodes under any ontology. $Sim(c_i, c_j)$ denotes the SS calculator for concepts c_i and c_j. $Sim(c_i, c_j) \in [0, 1]$. In our framework we consider relation R between any two concepts of the ontology belongs to either $hyponym \vee hypernym$.

Definitions

Definition 1. *$subsumers(c) = \{b \in V,\ c \in V \mid c \preceq b\} \cup \{c\}$, $c \preceq b$ signifies that concept c is a hierarchical specialization of concept b and “$\preceq$” represents the “subsumed by” relationship. $|subsumers(c)|$ denotes the number of the members of this set.*

Definition 2. *$hyponyms(c) = \{b \mid b \in V \wedge \forall\ b,\ b \preceq c\ \}$, i.e. set of concepts that are descendants of the concept c. For example, according to Figure 1, $hyponyms(Disorder\ of\ Thorax)$*

[1] https://wordnet.princeton.edu

[2] http://www.ihtsdo.org/snomed-ct

includes concepts like $Congestive\ Heart\ Failure$, Accute $Congestive\ Heart\ Failure$ and Biventricular $Congestive\ Heart\ Failure$. In our paper, we denote the set hyponym as $hypo(c)$. $|hypo(c)|$ denotes number of the members of this set.

Definition 3. $hypernyms(c) = \{b \mid b \in V \wedge \forall\ b,\ c \preceq b\ \}$, *i.e. the set of concepts that are ancestors of the concept c. For example, according to Figure 1, hypernyms(Rheumatoid Arthritis) includes Autoimmune Disease, $Disorder\ By\ Body\ Site$, and $SNOMEDRT + CTV3$. In this paper, we denote the set hypernym as $hyper(c)$. $|hyper(c)|$ denotes number of the members of this set.*

Definition 4. *instance hyponyms(c) = A set of concepts which are the real world entity of a concept c, e.g. instance hyponym of the "Planet" is "Earth".*

Definition 5. *depth(c) = The minimum distance between c and the root node of an ontology.*

Definition 6. $leaves(c) = \{l \in V,\ c \in V \mid l \in hyponyms(c)\ \Lambda\ l\ is\ a\ leaf\}$*, where, l is a leaf iff $hyponyms(l) = \phi\ \Lambda\ instance\ hyponyms(l) = \phi$. $|leaves(c)|$ denotes number of the members of this set.*

Definition 7. *nmih(c) = A set of subsumers that has direct relationship with the concept c by hyponym-hypernym relationship, e.g. according to Figure 1 $nmih(Pulmonary\ Edema) = \{Disorder\ Of\ Body\ System,\ Disorder\ of\ Thorax,\ Viscus\ Structure\ Finding\}$. $|nmih(c)|$ denotes number of the members of this set.*

Definition 8. *$Multiple\ Inheritance$ = When a concept c has multiple direct subsumers, then it has multiple inheritances.*

Definition 9. *$node_{max}$ = It represents maximum number of concepts of an ontology.*

Definition 10. *max_{leaves} = It denotes the number of leaf nodes the root node of an ontology has.*

Definition 11. *$depth_{max}$ = It is the maximum value among all the depth(c) for all concepts c belongs to the ontology.*

Related Work

Intrinsic IC calculation

We discuss some of the significant state of the art intrinsic IC calculation techniques in this section.

Several authors (Sánchez & Batet, 2012; Seco et al., 2004; Sánchez et al., 2011; Zhou et al., 2008; Meng et al., 2012; Adhikari et al., 2015) have proposed intrinsic IC calculation measures to overcome the limitations of corpora based IC calculation measures, discussed in "Introduction" section. According to this technique, it is calculated solely depending on the ontology itself. Seco et al. (Seco et al., 2004) is the pioneer in intrinsic way of calculating IC. Their model depends on the number of the hyponyms of the concept whose IC is to be calculated as shown below:

$$IC_{seco}(c_i) = \frac{\log(\frac{|hypo(c_i)|+1}{node_{max}})}{\log(\frac{1}{node_{max}})} \tag{1}$$

The main limitation of their model is, concepts that have equal number of hyponyms and different generality, will have same similarity score. To overcome this issue, Zhou et al. (Zhou et al., 2008) have introduced the idea of *depth*(c) with the *hyponyms*(c) as follows:

$$IC_{zhou}(c_i) = K(1 - \frac{\log(|hypo(c_i)| + 1)}{\log(node_{max})}) + (1 - K)\frac{\log(depth(c_i))}{\log(depth_{max})} \tag{2}$$

However, in their model, *depth*(c) has to be empirically tuned. Sánchez et al. (Sánchez et al., 2011) have overcome this problem and proposed a new intrinsic IC calculator, as follows:

$$IC_{sanchez}(c_i) = -\log(\frac{\frac{|leaves(c_i)|}{|subsumers(c_i)|} + 1}{max_{leaves} + 1}) \tag{3}$$

Besides the above, there are some other issues involved in the existing approach. For instance, according to Sánchez et al. (Sánchez et al., 2011) when concepts have the equal number of subsumers and leaves but different topological orientation of hyponyms with different number of hyponyms, then the IC calculator produces same IC values. This means the concepts have the same meaning but in reality, they carry different information. Meng et al. (Meng et al., 2012) have overcome this issue by considering the depth of the concept and depth of each hyponym of the concept as follows:

$$IC_{meng}(c_i) = \frac{log(depth(c_i))}{log(depth_{max})} \times (1 - \frac{log((\sum_{a \in hypo(c_i)} \frac{1}{depth(a)}) + 1)}{log(node_{max})}) \tag{4}$$

Although, in Meng et al. IC model, there is a possibility that two concepts have same hyponym structure, stay in the same depth, but have a different number of subsumers. In that case, their IC model will generate same IC value for both the concepts, which is not expected as the topological structure of both the concepts in the ontology is different.

Sánchez and Batet (Sánchez & Batet, 2012) have proposed an IC calculation model determining the commonness of a concept as follows:

$$IC_{s\&b}(c_i) = -\log(\frac{commonness(c_i)}{commonness(root)}) \tag{5}$$

where, $commonness(c_i) = \Sigma commonness(l), \forall\, l \mid l$ is a leaf node Λ l is subsumed by concept c_i and c_i is not a leaf node. $commonness(l) = \frac{1}{|subsumers(l)|}$. This IC model relies only on the number of subsumers of leaves of the concept whose IC is going to be calculated. But it should not be the only criteria to be considered in calculating IC.

To address the above mentioned issues present in $IC_{meng}(c_i)$ and $IC_{s\&b}(c_i)$ in our previous work (Adhikari et al., 2015), we have proposed an intrinsic IC calculation model as follows:

$$\begin{aligned} IC_{adhikari}(c) = {} & \frac{\log(depth(c) + 1)}{\log(depth_{max} + 1)} \\ & \times (1 - \log(\frac{\frac{|leaves(c)| \times (|nmih(c)|)}{max_{leaves}}}{|subsumers(c)|} + 1)) \\ & \times (1 - \frac{\log((\sum_{a \in hypo(c)} \frac{1}{depth(a)}) + 1)}{\log(node_{max})}) \end{aligned} \tag{6}$$

Yuan et al. (Yuan, Yu, & Wang, 2013) calculates IC intrinsically using some different topological factors in the following way:

$$IC_{yuan}(c_i) = f_{depth}(c_i) * (1 - f_{leaves}(c_i)) + f_{hypernyms}(c_i) \tag{7}$$

where, $f_{depth}(c_i)$, $f_{leaves}(c_i)$, $f_{hypernyms}(c_i)$ are defined as follows:

$$f_{depth}(c_i) = \frac{\log(depth(c_i))}{\log(depth_{max})} \tag{8}$$

$$f_{leaves}(c_i) = \frac{\log(|leaves(c_i)| + 1)}{\log(max_{leaves} + 1)} \tag{9}$$

$$f_{hypernyms}(c_i) = \frac{\log(|hyper(c_i)| + 1)}{\log(node_{max})} \tag{10}$$

where, c_i is a concept in an ontology. In this IC model the key factor that makes some improvement in finding IC of a concept is $f_{hypernyms}(c_i)$.

Semantic similarity models based on IC

Though several SS measures (Harispe et al., 2013) grounded in different theoretical bases exist, we consider only information-theoretic SS measures in this paper. There are basically IC based three classical SS models available as follows: First IC based SS model is proposed by Resnik (Resnik, 1995). Resnik defines a function for finding similarity between two concepts based upon IC of their Least Common Subsumer (LCS) (i.e. the common subsumer of the two concepts and has maximum IC) (Sánchez & Batet, 2011) in the following way:

$$Sim_{res}(c_i, c_j) = IC(LCS(c_i, c_j)) \tag{11}$$

Resnik's model has some limitations like, concept pairs which have the same LCS, posses same similarity value. To solve this issue, Lin (Lin, 1998), Jiang and Conrath (Jiang & Conrath, 1997) have proposed their own models. Lin extended Resnik's similarity formula by considering ratio with summation of individual IC of each concept:

$$Sim_{lin}(c_i, c_j) = \frac{2 \times Sim_{res}(c_i, c_j)}{IC(c_i) + IC(c_j)} \tag{12}$$

Jiang and Conrath have proposed a new measure for finding semantic distance between two concepts in terms of IC as follows:

$$Dist_{j\&c}(c_i, c_j) = IC(c_i) + IC(c_j) - 2 \times Sim_{res}(c_i, c_j) \tag{13}$$

Apart from these three classical IC based SS models, Pirró and Euzenat (Pirró & Euzenat, 2010), Sánchez and Batet (Sánchez & Batet, 2011) also have proposed IC based SS calculators. Pirró has presented a framework, which maps the feature-based model of SS into the information theoretic domain (Pirró, 2009). Pirró and Euzenat have extended that work and proposed a new FaITH model (Pirró & Euzenat, 2010) as follows:

$$Sim_{FaITH}(c_i, c_j) = \frac{IC(LCS(c_i, c_j))}{IC(c_i) + IC(c_j) - IC(LCS(c_i, c_j))} \tag{14}$$

In this similarity model, Pirró and Euzenat have used the IC model originally designed by Seco et al. (equation 1). Sánchez and Batet (Sánchez & Batet, 2011) mapped an edge counting SS measure into an IC based SS model:

$$\begin{aligned} Sim_{s\&b}(c_i, c_j) = \\ -\log \frac{IC(c_i) + IC(c_j) - 2 \times IC(LCS(c_i, c_j)) + 1}{2 \times max\_IC} \end{aligned} \tag{15}$$

In this model, Sánchez and Batet have used Sánchez's IC model (equation 3).

Proposed Approach

Before we introduce our proposed intrinsic IC based SS calculation model, we first discuss some of the limitations of LCS, the core of the existing state of the art intrinsic IC based similarity calculation models.

Limitations of LCS based SS model

LCS is a common subsumer of concepts c_i and c_j and has maximum IC among all the other common subsumers of c_i and c_j. The IC value of such common subsumer for any two concepts c_i and c_j is unique in a particular ontology. Thus, we ignore some other IC values of some special common subsumers of c_i and c_j. It is evident from the related work section that finding similarity between two concepts in an information theoretic way mainly relies on IC values. So, we are talking about some special common subsumers whose IC value matters for deciding similarity but are not considered in LCS based similarity models. Such set of common subsumers has some special criteria as follows: (i) no one is connected to each other by any hyponym-hypernym relationship, and (ii) the first member of such common subsumer set stays at maximum depth among all the common subsumers. To become next member of such common subsumer set, it has to stay at maximum depth among rest of the common subsumers and in this way all the members of such special set of common subsumers are selected provided that no one is in the relationship of hyponym-hypernym to each other. By the nature of their graphical orientation, those members of subsumer set become the most specific in their own conceptual dimension as a common subsumers of concepts c_i and c_j because the IC value of any node is directly proportional to its depth and it increases as we proceed towards the leaf nodes (Sánchez et al., 2011) and the more the IC value is, the more specific concept it becomes. Hence, these common subsumers introduce distinct dimensions to the concepts c_i and c_j while deciding similarity. Several such kind of common subsumers are available in any ontology. For instance, Autoimmune Disease and Disorder By Body Site are such common subsumers of Rheumatoid Arthritis and Lupus Erythematosus in SNOMED-CT ontology.

In the following, before we present our proposed SS model, we first illustrate its core idea $DCS(c_i, c_j)$ followed by the algorithm for finding $DCS(c_i, c_j)$.

What is DCS?

$DCS(c_i, c_j)$ stands for Disjoint Common Subsumers of c_i and c_j. It represents a set of nodes which are subsumers of both c_i and c_j but no one is in the relationship of hyponym-hypernym to each other. The first member of this set stays at maximum depth among all the common subsumers of c_i and c_j, and rest of the members of the set are selected by picking every time the deepest common subsumer which preserves the property of not having hyponym-hypernym relationship to each other. dcs represents the members of the set $DCS(c_i, c_j)$.

Each dcs acts as a distinct common subsumer having no relationship to other dcs. What we mean by the members of the set $DCS(c_i, c_j)$ stay at their maximum possible depth is explained as follows: suppose there is another node called dcs_x which is a parent node of dcs_j. Let depth of dcs_x and dcs_j are y and z respectively such that value of z is greater than value of y. Suppose, dcs_x also is a common subsumer of concepts c_i and c_j and has no hyponym-hypernym relationship with dcs_i and dcs_k. But we will consider dcs_j over dcs_x as a member of the set $DCS(c_i, c_j)$ because dcs_j stays at maximum depth preserving the property that no one is connected to each other by hyponym-hypernym relationship.

A real snippet from SNOMED-CT ontology is shown in *Figure 1*. In this figure all the circles represent concepts. Root node of SNOMED-CT is SNOMED RT+CTV3. Circle with bold edge denotes the possible members of set $DCS(c_i, c_j)$. In this snippet, concepts Congestive Heart Failure and Pulmonary Edema have three dcs, such as, Disorder Of Body System, Disorder Of Thorax, and Viscus Structure Finding. Further to note, concepts Rheumatoid Arthritis and Lupus Erythematosus have two dcs, such as, Autoimmune Disease and Disorder By Body Site.

Algorithm for finding $DCS(c_i, c_j)$

To find out the $DCS(c_i, c_j)$, earlier we have proposed an algorithm (Adhikari et al., 2016) as discussed below. The procedure for calculating $DCS(c_i, c_j)$ is provided in Algorithm 1. According to this algorithm, in Step 2, we first consider four empty sets: $DCS(c_i, c_j)$, $DCS_{suspect}(c_i, c_j)$, S_i, and S_j. In Step 3, store all the subsumers of concept c_i and c_j in S_i and S_j respectively. In Step 4, find the intersection of these two sets S_i and S_j and store them in $DCS_{suspect}(c_i, c_j)$. In Step 5, perform a sorting operation on set $DCS_{suspect}$ (c_i, c_j) in descending order based on the depth of each element of that set. In Step 6, pick the largest element as first dcs from the set $DCS_{suspect}(c_i, c_j)$ and assign to $DCS(c_i, c_j)$ and in Step 7, discard that element from $DCS_{suspect}(c_i, c_j)$. In Step 8, do the following operations until all the elements of $DCS_{suspect}(c_i, c_j)$ are checked: pick the next largest element x from $DCS_{suspect}(c_i, c_j)$ and check whether any of the element of $DCS(c_i, c_j)$ is hyponym of x. If none of the elements of $DCS(c_i, c_j)$ are the hyponym of x then add this x into set $DCS(c_i, c_j)$ and repeat Step 7. Otherwise discard x from $DCS_{suspect}(c_i, c_j)$ only, and do not add to $DCS(c_i, c_j)$. Step 9 ends the process.

Proposed IC based Semantic Similarity Model

In designing our similarity calculator, we consider pure information theoretic perspective and the structural aspect, i.e. $DCS(c_i, c_j)$ of the underlying ontology. Besides

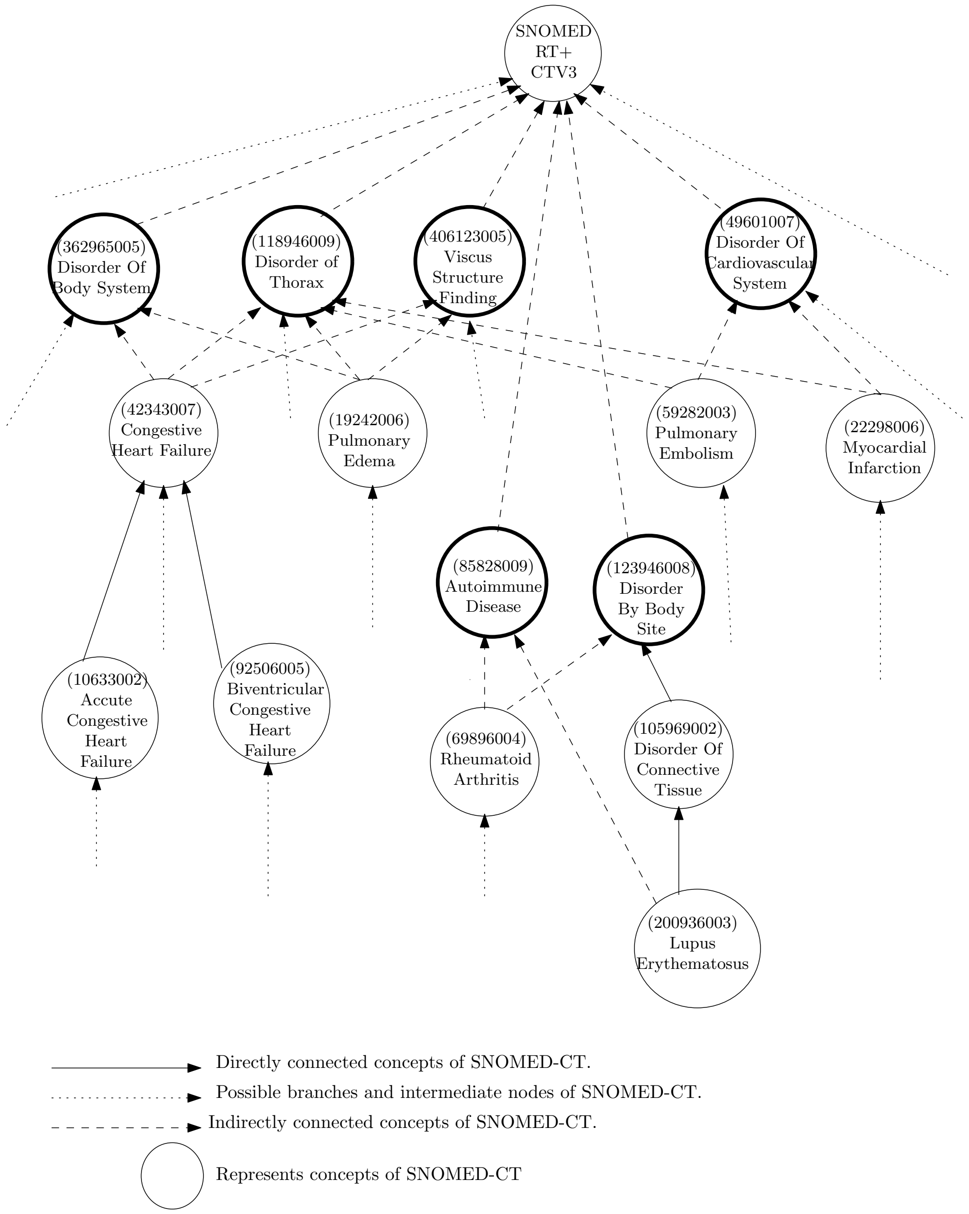
SNOMED RT+ CTV3
(362965005) Disorder Of Body System
(118946009) Disorder of Thorax
(406123005) Viscus Structure Finding
(49601007) Disorder Of Cardiovascular System
(42343007) Congestive Heart Failure
(19242006) Pulmonary Edema
(59282003) Pulmonary Embolism
(22298006) Myocardial Infarction
(85828009) Autoimmune Disease
(123946008) Disorder By Body Site
(10633002) Accute Congestive Heart Failure
(92506005) Biventricular Congestive Heart Failure
(69896004) Rheumatoid Arthritis
(105969002) Disorder Of Connective Tissue
(200936003) Lupus Erythematosus
Directly connected concepts of SNOMED-CT.
Possible branches and intermediate nodes of SNOMED-CT.
Indirectly connected concepts of SNOMED-CT.
Represents concepts of SNOMED-CT

Figure 1. A snippet from SNOMED-CT ontology with several *dcs*.

Algorithm 1 Calculating $DCS(c_i, c_j)$

```
Step 1. Start
Step 2. Initialize set $DCS(c_i, c_j) \leftarrow \phi$, set $DCS_{suspect}(c_i, c_j) \leftarrow \phi$, set $S_i \leftarrow \phi$, set $S_j \leftarrow \phi$.
Step 3. For concept $c_i$ add all subsumers($c_i$) to set $S_i$ and for concept $c_j$ add all subsumers($c_j$) to set $S_j$.
Step 4. set $DCS_{suspect}(c_i, c_j) \leftarrow S_i \cap S_j$.
Step 5. Sort elements of set $DCS_{suspect}(c_i, c_j)$ in descending order based on depth of each elements.
Step 6. Assign the largest element of the set $DCS_{suspect}(c_i, c_j)$ as first $dcs$ to $DCS(c_i, c_j)$.
Step 7. set $DCS_{suspect}(c_i, c_j) \leftarrow DCS_{suspect}(c_i, c_j)$ - $DCS(c_i, c_j)$.
Step 8. for each element $x$ in $DCS_{suspect}(c_i, c_j)$
          if any member of $DCS(c_i, c_j)$ is in the set $hyponyms(x)$
            discard that node $x$ from $DCS_{suspect}(c_i, c_j)$
          else add $x$ to $DCS(c_i, c_j)$ and repeat Step 7.
Step 9. End
```

$DCS(c_i, c_j)$, we also consider some ratio factors for formulating our model as follows:

$$Sim_{our}(c_i, c_j) = \frac{\sum_{r=1}^{m}\{\frac{IC(dcs_r)}{IC(c_i)-(IC(dcs_r)\times k)+1} + \frac{IC(dcs_r)}{IC(c_j)-(IC(dcs_r)\times k)+1}\}}{m} \tag{16}$$

where, m denotes size of the set $DCS(c_i, c_j)$ for concepts c_i and c_j. We focus on ontology with hyponym-hypernym (is-a) relationship and in such organization of concepts in a particular path from the root node, the deepest node (concept) should be the most specific (is-a type) concept (Sánchez et al., 2011). In our DCS finding algorithm we segregate several distinct paths generated from the root node where the deepest node acts as the member of the $DCS(c_i, c_j)$ set. Each member of $DCS(c_i, c_j)$ set actually holds the distinct semantics which act as an influencing factor for any two concepts c_i and c_j because, no one of the set $DCS(c_i, c_j)$ is connected to each other by any kind of conceptual relationship. For instance, Congestive Heart Failure is-a type of Disorder Of Body System, Disorder Of Thorax, and also is-a type of Viscus Structure Finding. Hence Congestive Heart Failure is influenced by three different senses. Another concept like Pulmonary Edema is also is-a type of Disorder Of Body System, Disorder Of Thorax, and also is-a type of Viscus Structure Finding. All of the later three concepts are members of the set $DCS(\mathit{Congestive\ Heart\ Failure},\ \mathit{Pulmonary\ Edema})$ and no one is connected to each other by any conceptual relationship. So, when two concepts are having more number of such distinct influencing subsumers in common, the chances for becoming more semantically similar also increases. Further to note that, the far any concept c is from its dcs_r, the more the IC value it has and becomes a factor for reducing the overall SS by producing a larger denominator in the summation part of our similarity model. In our similarity model we subtract IC of each dcs_r [for r = 1 to m] from IC of individual concept c_i and c_j in the denominator of the summation part. A concept c_i or c_j can be connected to multiple dcs_r inheriting several unique conceptual dimensions. So, by subtracting IC of each dcs_r from IC of concept c_i and c_j, we try to extract the amount of influence created by the other members of the set $DCS(c_i, c_j)$. Then we find out the degree of influence of such other members of the set $DCS(c_i, c_j)$ to each concepts c_i and c_j with respect to the influence of each dcs_r and summing up them all in the numerator part of our similarity model.

We use a binary parameter (k = 0 or 1) to account the influence of the members of $DCS(c_i, c_j)$ in the denominator of the summation part of our similarity measure. The reason behind considering such parameter is the structural differences (Pirró & Euzenat, 2010) that exist between different ontologies due to variable perceptions of each ontologists to fulfill their purposes. Based on our research observation on different ontologies (e.g. WordNet, SNOMED-CT, MeSH etc.), we have found some key ontological differences between the ontologies. These structural differences are the important factors which influence the designing of a generic similarity model. One of the such differences is the multiple number of members of the set $DCS(c_i, c_j)$ for two concepts. In some ontologies, some structural aspects like multiple number of dcs_r present in a larger magnitude, whereas in some ontologies the presence is negligible or missing. Hence, in the case where such structural aspect is negligible, the consideration of that aspect as a factor in designing similarity model is inappropriate. Hence, to account the influence of such structural factor we use a binary parameter k (where k = 0 or 1) in our SS model. We consider k = 0 when the presence of the structural aspect in an ontology is negligible. For instance, in case of WordNet there is no multiple members in the $DCS(c_i, c_j)$ set (100% singleton $DCS(c_i, c_j)$ sets for M&C dataset and 99.9% singleton sets for Wordsim similarity gold standard dataset). So, in this case, k = 0. Whereas, we have found significant number of non-singleton $DCS(c_i, c_j)$ sets in SNOMED-CT (only 79.31% singleton $DCS(c_i, c_j)$ sets) and MeSH (only 92.30% singleton $DCS(c_i, c_j)$ sets) ontology. Hence we use K = 1 to account the influence of multiple members of the $DCS(c_i, c_j)$ set in deciding the accuracy of SS between concepts.

Note that, in some ontologies, for instance WordNet, some words have more than one senses, i.e. $W = \{s_1, s_2, s_3, \ldots, s_n\}$. Such words are termed as polysemic words (Freihat, Giunchiglia, & Dutta, 2013). In such cases, we compute SS as follows:

$$Sim(W_x, W_y) = \max\{sim_{\forall i,j}(s_{x_i}, s_{y_j})\} \tag{17}$$

where, s_{x_i}, s_{y_j} are i^{th} and j^{th} senses (i.e. concepts or synsets in the WordNet ontology) of polysemic words W_x and W_y respectively.

Experiments

This section describes the experimental set-up and the results from a detailed evaluation of our proposed intrinsic IC based information theoretic SS model i.e. $Sim_{our}(c_i, c_j)$ over the existing state of the art IC based similarity models.

Task Description

For a fair comparison, we implement all the competing intrinsic IC based similarity models enlisted in Table 1. Our proposed SS calculator has been evaluated pairing up with several state of the art IC models, such as Seco et al., Zhou et al., Sánchez et al. (Sánchez et al., 2011), Sánchez et al. (Sánchez & Batet, 2012), Meng et al., Yuan et al. and Adhikari et al.

It is worth mentioning here that to implement the Jiang and Conrath similarity model, we use the following transformed similarity function (equation 18) originally proposed by Seco et al. (Seco et al., 2004).

$$Sim(c_i, c_j) = 1 - (\frac{IC(c_i) + IC(c_j) - 2 \times Sim_{res}(c_i, c_j)}{2}) \tag{18}$$

In our experiment, we consider three different ontologies. WordNet version 3.0 is chosen as a generic ontology. Beside these two generic ontologies, we also consider another two domain specific ontologies from the biomedical domain, namely, SNOMED-CT version May 2016 and MeSH version Rh–mesh 2014.

We use Pearson's correlation coefficient (CC) (Harispe, Sánchez, Ranwez, Janaqi, & Montmain, 2014; Pirró & Euzenat, 2010) as the metric to evaluate the accuracy of our proposed IC based similarity model. Note that, all the CC values are rounded up to two decimal points.

Table 1
Shows the competing IC based similarity models used in our paper.

Approach
Resnik (Resnik, 1995)
Lin (Lin, 1998)
Jiang and Conrath (Jiang & Conrath, 1997)
Sánchez and Batet (Sánchez & Batet, 2011)
Pirró and Euzenat (Pirró & Euzenat, 2010)

Experimental Setup

Datasets. For general benchmark dataset, we use Miller and Charles' (Miller & Charles, 1991) benchmark dataset (M&C dataset) and Wordsim similarity goldstandard (Sánchez & Batet, 2012) benchmark dataset. Wordsim similarity goldstandard is a subset of the WordSim353 [3] [4] dataset. It has a set of 203 word-pairs. But only 201 noun pairs from that Wordsim dataset are selected for our experiment.

For domain-specific benchmark dataset, we use Pedersen et al. 2007 dataset [5]. To derive a reliable test set, Pedersen et al. create this dataset with the help of three physicians and 9 medical coders. In original dataset, there are 30 term pairs. The pair "Chronic obstructive pulmonary disease" and "lung infiltrates" has been excluded from the original test bed because the term "lung infiltrates" is not found in the SNOMED-CT (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). Thus, the resulting test set consists of 29 pairs.

For MeSH, we use a subset of Pedersen et al. 2007 dataset. This subset consists of 26 term pairs. Of them, we remove four-term pairs such as Stomach cramps, Lung infiltrates, Rectal polyp and Entire Knee meniscus dataset due to their non-availability in MeSH ontology. It is worth to mention here that MeSH ontology has multiple roots. Even among the 26 term pairs of Pedersen dataset, there are 18 term pairs which have no root in common. To resolve this issue, we introduce a hypothetical root which acts as a common root to all the existing roots of MeSH ontology. Thus we have a single common root to MeSH ontology.

Ontology parsing libraries. We use python as the implementation language and several python based ontology parsing libraries. NLTK (Natural Language Toolkit) [6] is

[3] http://www.cs.technion.ac.il/ gabr/resources/data/wordsim353/

[4] http://www.semantic-measures-library.org/sml/

[5] http://www.semantic-measures-library.org/sml/

[6] http://www.nltk.org/

used to handle WordNet version 3.0.

For Mesh ontology, we use Pronto [7]. It is a Python library designed to work with ontologies. It can parse obo and owl/xml formats, open ontologies on the local host or from a network location.

For SNOMED-CT, we use PyMedTermino (Medical Terminologies for Python) [8] which is a Python module for easy access to the main medical terminologies. PyMedTermino is a product of the LIMICS research lab of Paris 13 University.

Results

We analyze and evaluate our proposed SS model over the existing state of the art intrinsic IC based SS models. For this evaluation, we implement all the competing methods enlisted in Table 1. For evaluating our proposed model, we use the above mentioned three ontologies such as: WordNet, SNOMED-CT, MeSH. Several benchmark datasets are also considered for comparison between our measure and state of the art measures. The details of results and evaluation are described as follows.

(i) Experiment on WordNet ontology

For WordNet ontology we choose two datasets to perform our experiment. 30 noun pairs from M&C dataset and 201 noun pairs from Wordsim similarity goldstandard.

Table 2 shows the results gained by our proposed model when 30 noun pairs of M&C dataset are selected. Based on the experimental results shown in Table 2 it is evident that our SS calculator with IC_{meng} shows significant result than most of the existing IC based similarity models giving CC of 0.87 with $M\&C$ dataset. This CC value is close enough to the upper limit for 30 noun pairs which is 0.88 (Resnik, 1995). Correlation value obtained by our SS model based on $IC_{adhikari}$ (correlation=0.86) is also very close to the upper bound.

It is clear from the Table 2 that our SS calculator embedded with the existing IC calculators can perform efficiently as compared to most of the existing methods when WordNet ontology and 30 noun pairs of $M\&C$ dataset are choosen.

To reaffirm the accuracy of our similarity model in WordNet ontology, we evaluate it with a bigger dataset. We evaluate our proposed similarity model more precisely using a recent generic benchmark dataset, namely Wordsim similarity goldstandard consisting of 201 noun pairs from WordNet ontology. Table 3 shows the correlation values for this new dataset. From Table 3, it is evident that our SS model gives a significant correlation with human evaluation over the existing SS measures. Our similarity model embedded with IC_{yuan} and $IC_{adhikari}$ produces very high correlation over the existing similarity models. It is also noticeable that all the correlation scores generated by our similarity model are very close to each other.

(iii) Experiment on SNOMED-CT ontology

Based on Pedersen et al. dataset, we find the correlation with physicians, medical coders, and an average of both of them but we measure the efficiency of our proposed similarity

[7] http://pronto.readthedocs.io/en/latest/

[8] http://pythonhosted.org/PyMedTermino/

Table 2
Shows CC values of different existing intrinsic IC based similarity models and our similarity model for 30 noun-pairs of M&C dataset.

Similarity models	IC model	Correlation with M&C
Resnik	(IC_{seco})	0.80
Resnik	(IC_{zhou})	0.83
Resnik	$(IC_{sanchez})$	0.83
Resnik	$(IC_{s\&b})$	0.81
Resnik	(IC_{meng})	0.86
Resnik	(IC_{yuan})	0.83
Resnik	$(IC_{adhikari})$	0.86
Lin	(IC_{seco})	0.84
Lin	(IC_{zhou})	0.82
Lin	$(IC_{sanchez})$	0.84
Lin	$(IC_{s\&b})$	0.84
Lin	(IC_{meng})	0.86
Lin	(IC_{yuan})	0.84
Lin	$(IC_{adhikari})$	0.86
Jiang and Conrath	(IC_{seco})	0.88
Jiang and Conrath	(IC_{zhou})	0.82
Jiang and Conrath	$(IC_{sanchez})$	0.87
Jiang and Conrath	$(IC_{s\&b})$	0.88
Jiang and Conrath	(IC_{meng})	0.83
Jiang and Conrath	(IC_{yuan})	0.82
Jiang and Conrath	$(IC_{adhikari})$	0.84
Sánchez and Batet	(IC_{seco})	0.88
Sánchez and Batet	(IC_{zhou})	0.84
Sánchez and Batet	$(IC_{sanchez})$	0.86
Sánchez and Batet	$(IC_{s\&b})$	0.85
Sánchez and Batet	(IC_{meng})	0.85
Sánchez and Batet	(IC_{yuan})	0.84
Sánchez and Batet	$(IC_{adhikari})$	0.85
Pirró and Euzenat	(IC_{seco})	0.83
Pirró and Euzenat	(IC_{zhou})	0.85
Pirró and Euzenat	$(IC_{sanchez})$	0.85
Pirró and Euzenat	$(IC_{s\&b})$	0.83
Pirró and Euzenat	(IC_{meng})	0.86
Pirró and Euzenat	(IC_{yuan})	0.84
Pirró and Euzenat	$(IC_{adhikari})$	0.86
Our_Sim	(IC_{seco})	0.83
Our_Sim	(IC_{zhou})	0.85
Our_Sim	$(IC_{sanchez})$	0.84
Our_Sim	$(IC_{s\&b})$	0.84
Our_Sim	(IC_{meng})	0.87
Our_Sim	(IC_{yuan})	0.84
Our_Sim	$(IC_{adhikari})$	0.86

Table 3
Shows CC values of different existing intrinsic SS models and our proposed similarity model for 201 noun-pairs of Wordsim similarity goldstandard dataset.

Similarity models	IC model	Correlation with M&C
Resnik	(IC_{seco})	0.66
Resnik	(IC_{zhou})	0.64
Resnik	$(IC_{sanchez})$	0.66
Resnik	$(IC_{s\&b})$	0.67
Resnik	(IC_{meng})	0.67
Resnik	(IC_{yuan})	0.68
Resnik	$(IC_{adhikari})$	0.68
Lin	(IC_{seco})	0.69
Lin	(IC_{zhou})	0.64
Lin	$(IC_{sanchez})$	0.66
Lin	$(IC_{s\&b})$	0.69
Lin	(IC_{meng})	0.68
Lin	(IC_{yuan})	0.69
Lin	$(IC_{adhikari})$	0.68
Jiang and Conrath	(IC_{seco})	0.67
Jiang and Conrath	(IC_{zhou})	0.63
Jiang and Conrath	$(IC_{sanchez})$	0.66
Jiang and Conrath	$(IC_{s\&b})$	0.67
Jiang and Conrath	(IC_{meng})	0.66
Jiang and Conrath	(IC_{yuan})	0.65
Jiang and Conrath	$(IC_{adhikari})$	0.66
Sánchez and Batet	(IC_{seco})	0.69
Sánchez and Batet	(IC_{zhou})	0.66
Sánchez and Batet	$(IC_{sanchez})$	0.68
Sánchez and Batet	$(IC_{s\&b})$	0.68
Sánchez and Batet	(IC_{meng})	0.68
Sánchez and Batet	(IC_{yuan})	0.67
Sánchez and Batet	$(IC_{adhikari})$	0.68
Pirró and Euzenat	(IC_{seco})	0.70
Pirró and Euzenat	(IC_{zhou})	0.68
Pirró and Euzenat	$(IC_{sanchez})$	0.69
Pirró and Euzenat	$(IC_{s\&b})$	0.70
Pirró and Euzenat	(IC_{meng})	0.71
Pirró and Euzenat	(IC_{yuan})	0.71
Pirró and Euzenat	$(IC_{adhikari})$	0.71
Our_Sim	(IC_{seco})	0.68
Our_Sim	(IC_{zhou})	0.65
Our_Sim	$(IC_{sanchez})$	0.66
Our_Sim	$(IC_{s\&b})$	0.68
Our_Sim	(IC_{meng})	0.68
Our_Sim	(IC_{yuan})	0.69
Our_Sim	$(IC_{adhikari})$	0.69

model based on finding the correlation with average similarity scores of physicians and medical coders. Table 4 shows the experimental results of our proposed model and comparison with the existing methodologies respectively when SNOMED-CT is considered as an ontology. Experimental results presented in Table 4 show our similarity calculator with $IC_{sanchez}$ surpasses all the existing SS measures. It produces highest correlation (0.76) with human evaluation than any of the state of the art SS models. Even our similarity model embedded with IC_{seco} and $IC_{s\&b}$ surpass state of the art similarity models.

(iv) Experiment on MeSH ontology

Here, we also measure the efficiency of our proposed framework based on finding the correlation with average similarity scores of physicians and medical coders. It is evident from results shown in Table 5 that our similarity calculator with the existing IC calculators gives high correlation values with human evaluation. Also, our SS calculator with $IC_{s\&b}$ surpasses all of the existing SS measures producing maximum correlation value (0.56) with human evaluation except $Sim_{FaITH}(c_i, c_j)$ embedded with IC_{meng} (0.57). Even our similarity model embedded with $IC_{sanchez}$ and IC_{meng} surpass most of the existing similarity models.

Discussion

In case of SNOMED-CT ontology, our similarity model gives the highest correlation 0.76, whereas state of the art similarity models produce maximum correlation value 0.69. The second best correlation score produced by the state of the art measures is 0.67 when we select SNOMED-CT ontology. It is worth to notice that the difference between these two correlation scores (i.e. 0.69 and 0.67) produced by the existing measures is 0.02. Whereas the difference between the best score produced by our model and the maximum score produced by the state of the art measures (i.e. 0.76 and 0.69) is 0.07 and the correlation scores, generated for all the ontologies in our experiment, ranges from 0 to 1. Hence, 0.07 improvement over the maximum score (i.e. 0.69) produced by the state of the art measures signifies the usefulness of our proposed model. We combine our proposed similarity model with the state of the art IC calculators mentioned in “Related Work” section and among them, our similarity model with IC_{seco}, $IC_{sanchez}$, and $IC_{s\&b}$ surpass all the existing state of the art thirty-five combinations of similarity models and IC calculators.

In MeSH ontology, our proposed model produces a very high correlation of 0.56 with human evaluation and no one among all the existing thirty-five combinations of similarity models and IC calculators surpass our result except $Sim_{FaITH}(c_i, c_j)$ embedded with IC_{meng}. All these results show that our proposed model gives similarity scores which are highly correlated with human evaluation as compared to the state of the art intrinsic IC based similarity models. Even our similarity model surpasses most of the (32 out of 35 combinations of similarity models and IC calculators) existing intrinsic IC based similarity model and IC calculator combinations when WordNet ontology is selected, producing a very high correlation 0.87.

Further to state that, there is no such fixed combination of state of the art similarity calculator and IC calculator which gives the best result for any experimental ontology. For instance, among all the existing methods, the combination of Pirró & Euzenat similarity model (equation 14) and IC_{seco} gives the best correlation scores when SNOMED-CT ontol-

Table 4
Shows CC values of different existing intrinsic IC based similarity models and Our similarity model for SNOMED-CT ontology.

Similarity models	IC models	Corr. Phy.	Corr. M.Co.	Corr. Avg.
Resnik	(IC_{seco})	0.56	0.56	0.58
Resnik	(IC_{zhou})	0.49	0.44	0.49
Resnik	$(IC_{sanchez})$	0.53	0.52	0.55
Resnik	$(IC_{s\&b})$	0.52	0.51	0.53
Resnik	(IC_{meng})	0.59	0.56	0.60
Resnik	(IC_{yuan})	0.54	0.52	0.55
Resnik	$(IC_{adhikari})$	0.54	0.52	0.55
Lin	(IC_{seco})	0.60	0.62	0.63
Lin	(IC_{zhou})	0.50	0.46	0.50
Lin	$(IC_{sanchez})$	0.55	0.56	0.58
Lin	$(IC_{s\&b})$	0.55	0.57	0.58
Lin	$(IC_{meng}\ )$	0.60	0.57	0.61
Lin	(IC_{yuan})	0.56	0.54	0.57
Lin	$(IC_{adhikari})$	0.56	0.54	0.57
Jiang and Conrath	(IC_{seco})	0.53	0.56	0.56
Jiang and Conrath	(IC_{zhou})	0.47	0.44	0.47
Jiang and Conrath	$(IC_{sanchez})$	0.53	0.54	0.56
Jiang and Conrath	$(IC_{s\&b})$	0.52	0.58	0.55
Jiang and Conrath	(IC_{meng})	0.52	0.50	0.53
Jiang and Conrath	(IC_{yuan})	0.50	0.48	0.51
Jiang and Conrath	$(IC_{adhikari})$	0.50	0.48	0.51
Sánchez and Batet	(IC_{seco})	0.57	0.62	0.62
Sánchez and Batet	(IC_{zhou})	0.52	0.50	0.53
Sánchez and Batet	$(IC_{sanchez})$	0.63	0.69	0.69
Sánchez and Batet	$(IC_{s\&b})$	0.60	0.67	0.66
Sánchez and Batet	(IC_{meng})	0.56	0.55	0.57
Sánchez and Batet	(IC_{yuan})	0.55	0.55	0.57
Sánchez and Batet	$(IC_{adhikari})$	0.55	0.55	0.57
Pirró and Euzenat	(IC_{seco})	0.63	0.69	0.69
Pirró and Euzenat	(IC_{zhou})	0.57	0.55	0.59
Pirró and Euzenat	$(IC_{sanchez})$	0.62	0.66	0.67
Pirró and Euzenat	$(IC_{s\&b})$	0.60	0.65	0.65
Pirró and Euzenat	(IC_{meng})	0.64	0.64	0.67
Pirró and Euzenat	(IC_{yuan})	0.62	0.64	0.66
Pirró and Euzenat	$(IC_{adhikari})$	0.62	0.63	0.66
Our_Sim	(IC_{seco})	0.65	0.70	0.70
Our_Sim	(IC_{zhou})	0.57	0.56	0.59
Our_Sim	$(IC_{sanchez})$	0.67	0.78	0.76
Our_Sim	$(IC_{s\&b})$	0.66	0.77	0.74
Our_Sim	(IC_{meng})	0.63	0.65	0.67
Our_Sim	(IC_{yuan})	0.60	0.64	0.64
Our_Sim	$(IC_{adhikari})$	0.62	0.65	0.66

Table 5

Shows CC values of different existing intrinsic IC based similarity models and Our similarity model for MeSH ontology.

Similarity models	IC models	Corr. Phy.	Corr. M.Co.	Corr. Avg.
Resnik	(IC_{seco})	0.39	0.39	0.40
Resnik	(IC_{zhou})	0.40	0.41	0.42
Resnik	$(IC_{sanchez})$	0.38	0.37	0.39
Resnik	$(IC_{s\&b})$	0.38	0.38	0.40
Resnik	(IC_{meng})	0.47	0.52	0.51
Resnik	(IC_{yuan})	0.45	0.48	0.48
Resnik	$(IC_{adhikari})$	0.45	0.48	0.48
Lin	(IC_{zhou})	0.41	0.42	0.43
Lin	(IC_{seco})	0.39	0.40	0.41
Lin	$(IC_{sanchez})$	0.38	0.38	0.39
Lin	$(IC_{s\&b})$	0.40	0.39	0.41
Lin	(IC_{meng})	0.48	0.53	0.52
Lin	(IC_{yuan})	0.45	0.49	0.49
Lin	$(IC_{adhikari})$	0.45	0.48	0.48
Jiang and Conrath	(IC_{seco})	0.34	0.33	0.34
Jiang and Conrath	(IC_{zhou})	0.39	0.39	0.40
Jiang and Conrath	$(IC_{sanchez})$	0.36	0.36	0.37
Jiang and Conrath	$(IC_{s\&b})$	0.44	0.41	0.44
Jiang and Conrath	(IC_{meng})	0.49	0.52	0.52
Jiang and Conrath	(IC_{yuan})	0.48	0.49	0.50
Jiang and Conrath	$(IC_{adhikari})$	0.46	0.48	0.49
Sánchez and Batet	(IC_{seco})	0.38	0.39	0.40
Sánchez and Batet	(IC_{zhou})	0.43	0.44	0.45
Sánchez and Batet	$(IC_{sanchez})$	0.44	0.47	0.47
Sánchez and Batet	$(IC_{s\&b})$	0.49	0.51	0.52
Sánchez and Batet	(IC_{meng})	0.51	0.56	0.56
Sánchez and Batet	(IC_{yuan})	0.50	0.54	0.54
Sánchez and Batet	$(IC_{adhikari})$	0.49	0.53	0.53
Pirró and Euzenat	(IC_{seco})	0.44	0.46	0.47
Pirró and Euzenat	(IC_{zhou})	0.46	0.49	0.49
Pirró and Euzenat	$(IC_{sanchez})$	0.42	0.44	0.45
Pirró and Euzenat	$(IC_{s\&b})$	0.44	0.46	0.47
Pirró and Euzenat	(IC_{meng})	0.51	0.58	0.57
Pirró and Euzenat	(IC_{yuan})	0.49	0.56	0.55
Pirró and Euzenat	$(IC_{adhikari})$	0.50	0.56	0.55
Our_Sim	(IC_{seco})	0.43	0.45	0.46
Our_Sim	(IC_{zhou})	0.44	0.46	0.47
Our_Sim	$(IC_{sanchez})$	0.50	0.56	0.55
Our_Sim	$(IC_{s\&b})$	0.51	0.56	0.56
Our_Sim	(IC_{meng})	0.49	0.55	0.54
Our_Sim	(IC_{yuan})	0.48	0.53	0.52
Our_Sim	$(IC_{adhikari})$	0.47	0.51	0.51

ogy is selected. Whereas the combination of Jiang & Conrath similarity model and $IC_{s\&b}$ gives the best correlation scores when WordNet ontology with M&C dataset is selected among all the existing methods.

Hence, in spite of the significant accuracy obtained by our proposed similarity model, the combination of our similarity calculator and an existing IC calculator which gives the best result also varies from ontology to ontology. For instance, in case of SNOMED-CT ontology, our similarity calculator produces the best result when embedded with $IC_{sanchez}$ (Sánchez et al., 2011). On the other hand, in case of MeSH, our similarity calculator produces the best result when embedded with $IC_{s\&b}$. *Figure 2* highlights the accuracy of our proposed similarity calculator when embedded with different IC calculators for a particular ontology. From that *Figure 2*, we can also identify which IC calculator gives the best result when embedded with our proposed similarity model for a particular ontology.

Further to note that, the accuracy of all the state of the art similarity measures, reported in their respective papers, are evaluated on a single ontology (Resnik, 1995; Lin, 1998; Jiang & Conrath, 1997; Sánchez & Batet, 2011; Pirró & Euzenat, 2010). The reason behind this is the structural differences of different ontologies will not allow to design a generic similarity model without analyzing the key differences between the ontologies and tune it based on the structure of the ontology. For example, we have found the structure of $DCS(c_i, c_j)$ changes from ontology to ontology (e.g. WordNet, SNOMED-CT, MeSH) significantly as described earlier in “Proposed Approach”. Hence, to produce a generic similarity measure we consider a binary parameter (k = 0 or 1) to count the influence of the members of $DCS(c_i, c_j)$ in such ontology where it actually matters.

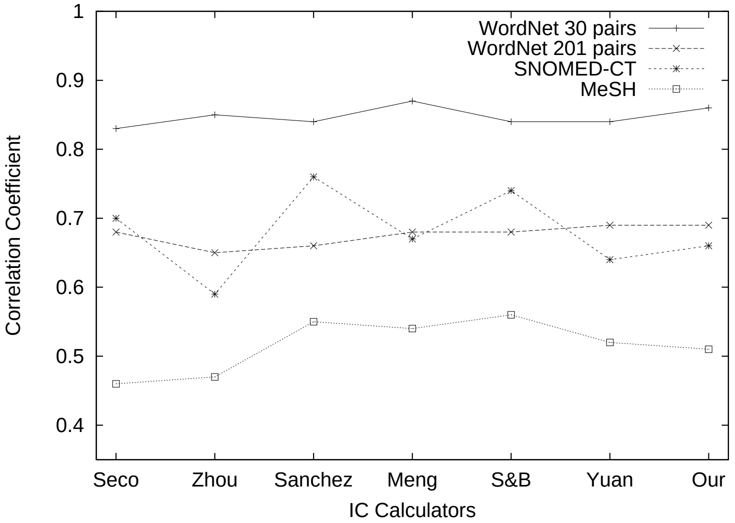

Figure 2. Shows the accuracy of our similarity calculator embedded with different IC calculators for different ontologies.

To further illustrate the accuracy of our proposed intrinsic IC based similarity model, we compare it with the state of the art corpora based IC dependent SS models such as, Resnik (Resnik, 1995), Lin (Lin, 1998), and J&C (Jiang & Conrath, 1997) similarity model. The motivation is to show the significance of intrinsic IC based similarity model over the corpora based measures. Resnik, Lin, and J&C IC calculation techniques are pure corpora based. It is not possible to simulate them in our experimental environment. Hence, we consider results obtained by their original simulations. For a fair comparison, we simulate our proposed model taking the same number of noun pairs from M&C dataset as they have considered. Correlation obtained by their methods range from 0.70 to 0.73 which is far lower than correlation obtained (0.86) by our proposed similarity model. Table 6 shows the comparison between corpora based measures vs our proposed model. The last column of the table shows the correlation values with respect to M&C dataset. It is clear from the Table 6 that intrinsic IC based similarity models give higher correlation compared to the corpora based IC dependent similarity models.

Table 6
Shows CC values of corpora based similarity models and our proposed similarity model for 28 term pairs.

Similarity models (with IC models)	Type	No. of noun pairs	Correlation with M&C
Resknik (Resnik IC (Resnik, 1995))	corpora-based	28	0.72
Lin (Resnik IC (Resnik, 1995))	corpora-based	28	0.70
Jiang and Conrath (Resnik IC (Resnik, 1995))	corpora-based	28	0.73
Our_Sim (Adhikari IC (Adhikari et al., 2015))	intrinsic	28	0.86

Conclusion

In this paper, we propose a domain-independent intrinsic IC based SS model for finding the similarity between concepts within a single ontology. In this work, we introduce a new structural aspect called $DCS(c_i, c_j)$. We experiment on three different ontologies ranging from generic (WordNet) to domain specific (SNOMED-CT, MeSH) ontologies. The experimental results show that our proposed SS calculation model is compatible enough to produce significant SS scores when embedded with the state of the art intrinsic IC calculators. $Table$ 6 shows that the proposed similarity calculation model also surpasses all the corpora based state of the art methods. Our proposed similarity model produces significant results when we conduct the experiments on MeSH and WordNet ontologies. Even the results gained from the experiments performed on ontology like SNOMED-CT show that our proposed similarity calculator surpasses all the existing state of the art approaches with a significant margin. The similarity value produced by our model is more accurate and effective because it achieves a high correlation with human evaluation compared to state of the art similarity models.

In the future, we aim to achieve the following goals. (i) Finding a combination of SS and IC calculator that produce the best result for any ontology. (ii) Design a SS calculator to find similarity between two concepts belonging to two or more than two distinct ontologies.

References

Adhikari, A., Singh, S., Dutta, A., & Dutta, B. (2015). A novel information theoretic approach for finding semantic similarity in wordnet. In *Tencon 2015-2015 ieee region 10 conference* (pp. 1–6).

Adhikari, A., Singh, S., Mondal, D., Dutta, B., & Dutta, A. (2016). A novel information theoretic framework for finding semantic similarity in wordnet. *CoRR*, *abs/1607.05422*. Retrieved from `http://arxiv.org/abs/1607.05422`

Aouicha, M. B., & Taieb, M. A. H. (2015). G2ws: Gloss-based wordnet and wiktionary semantic similarity measure. In *Computer systems and applications (aiccsa), 2015 ieee/acs 12th international conference of* (pp. 1–7).

Freihat, A. A., Giunchiglia, F., & Dutta, B. (2013). Approaching regular polysemy in wordnet. In *proceedings of 5th international conference on information, process, and knowledge management (eknow), nice, france.*

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013). Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base analysis. *arXiv preprint arXiv:1310.1285*.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of biomedical informatics*, *48*, 38–53.

Hliaoutakis, A. (2005). Semantic similarity measures in mesh ontology and their application to information retrieval on medline. *Master's thesis*.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Lin, D. (1998). An information-theoretic definition of similarity. In *Icml* (Vol. 98, pp. 296–304).

Meng, L., Gu, J., & Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in wordnet. *International Journal of Grid and Distributed Computing*, *5*(3), 81–94.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, *6*(1), 1–28.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, *40*(3), 288–299.

Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, *68*(11), 1289–1308.

Pirró, G., & Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. In *International semantic web conference* (pp. 615–630).

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of biomedical informatics*, *44*(5), 749–759.

Sánchez, D., & Batet, M. (2012). A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and*

Information Systems (IJSWIS), *8*(2), 34–50.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, *24*(2), 297–303.

Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th european conference on artificial intelligence* (pp. 1089–1090).

Yuan, Q., Yu, Z., & Wang, K. (2013). A new model of information content for measuring the semantic similarity between concepts. In *Cloud computing and big data (cloudcom-asia), 2013 international conference on* (pp. 141–146).

Zhou, Z., Wang, Y., & Gu, J. (2008). A new model of information content for semantic similarity in wordnet. In *Future generation communication and networking symposia, 2008. fgcns'08. second international conference on* (Vol. 3, pp. 85–89).