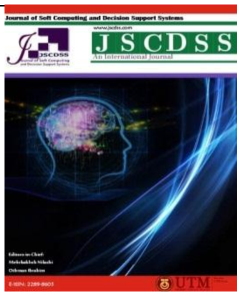
An International Journal
JSCDSS
http://www.jscdss.com
Vol.2 No.6 December 2015: 1-13
Article history:
Accepted 14 September 2015
Published online 14 September 2015

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

A Review of Semantic Similarity Measures in Biomedical Domain Using SNOMED-CT

Mojtaba Zare [a,*], Christina Pahl [a], Mehrbakhsh Nilashi [a], Naomie Salim [a], Othman Ibrahim [a]
[a] Faculty of Computing, Universiti Technologi Malaysia, Johor, Malaysia

*** Corresponding author email address**: mojtabazare123@yahoo.com

Abstract

The determination of semantic similarity between word pairs is an important task in text understanding that supports the processing, classification and structuring of textual resources. In the field of biomedical, semantic similarity measures have been the focus of much research by exploiting knowledge sources such as domain ontologies. SNOMED-CT as a main biomedical ontology provides a global and broad hierarchical terminology for clinical data storage, encoding, and the retrieval of health and diseases information. In this study, we classified the measures proposed in biomedical domain and used SNOMED-CT as an input ontology. We also examined the studies that evaluated these methods using biomedical benchmarks. Regarding this, three major databases, including Science Direct, Springer and IEEE were selected to extract studies which proposed similarity measures and used SNOMED-CT as a knowledge source. The purpose of this study is to provide the reader with the understanding about the application of semantic similarity measures in biomedical domain using SNOMED-CT, and to gain a clear insight about the performance of these methods. This study also supports researchers and practitioners in effectively adapting semantic similarity measures in SNOMED-CT and provides an insight into its state-of-the-art.

Keywords: Biomedical ontologies, SNOMED-CT, Semantic similarity measure

1. Introduction

Semantic Similarity Measures (SSMs) estimate the similarity between two given concepts (Janowicz et al., 2015; Liao et al., 2014; Sahni et al., 2014). The estimation of the semantic similarity between concepts helps in better understanding of textual resources (Song et al., 2014, Chakraborty et al., 2014). These measures are mainly categorized into two groups, including distributional based and knowledge based methods (Garla and Brandt, 2012a). Distributional based methods utilize the distribution of concepts within a corpus in conjunction with a knowledge source to compute similarity; these measures include corpus Information Content (IC) and context vector methods (Jiang et al., 2014). Knowledge based methods, on the other hand, utilize knowledge sources, such as ontologies and semantic networks. Knowledge based methods are divided into two groups, path-based and intrinsic IC-based measures (McInnes and Pedersen, 2015; Harispe et al., 2014).

Semantic similarity measures have been used in wide array of applications in biomedical domain, using biomedical ontologies. They have been applied to design information retrieval algorithms (Chaves-González and MartíNez-Gil, 2013; Uddin et al., 2013), to disambiguate texts (McInnes and Pedersen, 2013; Miller et al., 2012), to suggest drug repositioning (Gottlieb et al., 2011; Lamurias et al., 2013) and to cluster genes, according to their molecular function (Pesquita et al., 2009; Guzzi et al., 2012). Semantic similarity measures are indeed critical components of many knowledge-based systems (Chang and Lee, 2015; Gottlieb et al., 2011). In addition, they are nowadays receiving more attention due to the growing adoption of both Semantic Web and Linked Data paradigms (Bizer et al., 2009; Iosif and Potamianos 2015).

Semantic similarity measures, based on knowledge sources and ontologies, use the taxonomical evidences modeled in the ontology to assess the similarity of two given concepts. In fact, ontologies support these measures to model unstructured and heterogeneous information through the hierarchical vocabularies and structured sets of concepts (Harispe et al., 2014; Cross et al., 2013; Meng et al., 2013). Fortunately, the field of biomedical has been very prolific in creating medical ontologies which organize concepts in a non-ambiguous way to be used by semantic measures (Batet et al., 2011; SáNchez and Batet, 2011; Al-Mubaid and Nguyen 2009). Some well-known examples of ontologies in biomedical domain include Medical Subject

Headings (MeSH), International Classification of Diseases (the ICD taxonomy) and Systematised Nomenclature of Medicine, Clinical Terms (SNOMED-CT).

In this study, we reviewed and classified academic research that applied semantic similarity measures in biomedical domain using SNOMED CT as an input ontology. Regarding this, three major databases, including Science Direct, Springer and IEEE have been selected. The extracted papers from these databases have developed measures that used SNOMED-CT as a knowledge source. Hence, the objective of this study is twofold: First, to provide the reader with the understanding about the application of semantic similarity measures in biomedical domain using SNOMED-CT. Second, to explore biomedical studies that evaluated these measures based on biomedical benchmarks.

The reminder part of this manuscript is divided into the following sections: The definition of ontology and biomedical knowledge sources are presented in Section 2. In Section 3, an overview of different types of semantic similarity measures is presents. Section 4 focuses on summarizing and classifying the previous related works. In Section 5, biomedical reference standards are introduced. In Section 6, the semantic similarity evaluation studies are discussed. Finally, discussions and conclusions are presented in Sections 7 and 8, respectively.

2. Ontology

The term ontology is used in two different ways representing two different things. The first usage is “philosophical ontology” where “ontology is the study of being or existence”. In this definition, ontology comprises the basic subject of metaphysics that explains existence in a systematic manner. In order to have a systematic manner, philosophical ontology deals with the types and structures of objects, properties, events, processes and relations related to each part of reality (Zadeh and Reformat, 2013; Thomasson, 2014)). The second usage of ontology is “ontology and information systems” where ontologies represent relations among terms similar to taxonomies. But, in this field, the main difference between ontologies and taxonomies is that ontologies present richer and detailed meaning for the relationships among terms, attributes, and concepts in comparison with taxonomies (Sicilia, 2014; Zaid and Lau, 2014).

There exist numerous definitions of ontology. One of the earliest definition of ontology is that of Neches et al. (1991), stating that “An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary”. Gruber (1995) provided one of the most widely adopted definitions of ontology, as: “A formal, explicit specification of a shared conceptualization”. This definition emphasizes several important characteristics of an ontology (Studer et al., 1998).

“It is formal. This means that an ontology should be machine-readable.

It is explicit. It indicates that the type of concepts used in an ontology and the restrictions on their use are explicitly defined.

It is shared. It reflects the notion that the ontology captures consensual knowledge, which means, it is not the privilege of some individual, but accepted by a group.

It specifies a conceptualization. This refers to an abstract model of some phenomenon in the world by identifying relevant concepts of that phenomenon.”

2.1 Biomedical knowledge sources and ontologies – SNOMED CT and UMLS

Examples of ontologies in the biomedical domain include SNOMED-CT and MeSH. MeSH has been created for the purpose of indexing and is used for indexing articles from the Medline database, whereas the scope of SNOMED-CT in health care is specially to model clinical data, i.e., to assist in annotating Electronic Health Records (EHRs). UMLS, on the other hand, is a knowledge source, containing several biomedical ontologies and vocabularies such as MeSH and SNOMED-CT. In this study our focus is on the studies that performed in the domain of SNOMED-CT. Hence, we also consider studies conducted within the framework of UMLS, since UMLS consists of SNOMED-CT concepts. In this section we adhere to a broad understanding of SNOMED-CT and UMLS knowledge source.

2.1.1 SNOMED-CT

"SNOMED-CT" stands for "Systematized Nomenclature of Medicine — Clinical Terms" is a systematically organized computer readable collection of medical terminology covering most areas of clinical information where its first version was released in 2002. SNOMED-CT ontology provides a global and broad hierarchical terminology for clinical data storage, encoding, and the retrieval of health and diseases information (Lee et al., 2014; Schulz et al., 2014; Schulz and Martínez-Costa, 2013). Basically, SNOMED-CT has been designed to be used by computer applications to represent clinical data in consistent and unambiguous manner. Then, the resulted data can be used for electronic health records (EHRs) and decision-support (DS) systems and finally to enable semantic interoperability which is precisely the goal (Sicilia, 2014; Campbell et al., 2013, Duarte et al., 2014). SNOMED-CT as an internationally accepted standard ontology is included in the UMLS repository.

Regarding the structure of SNOMED-CT, its concepts are organized in a hierarchical structure to ease and permit searching concepts at various levels of specificity. In this ontology, concepts are connected by two main relations, including: parent-child and broader-narrower. The parent-child relation is strictly IS-A relation, but the broader-narrower relation contains part-of relation (McInnes and Pedersen, 2015).

An International Journal
JSCDSS
E-ISSN: 2289-8603

Fig. 1 shows the hierarchical structure of SNOMED-CT. The first level of this hierarchy contains 19 categories, ranging from body structure to physical object, followed by its second level where it comprises 345 categories and this structure continues further down until very specific concepts are reached. As can be seen from the figure, the root node of all concepts is “SNOMED-CT”.

Table 1 gives information about all 19 categories of hierarchy first-level in SNOMED-CT, listing in descending-size order. In this table, the size of a category is defined by the total number of concepts under each category.

As can be seen in Fig.1, the concept “a” can belong to multiple categories such as “Finding by site” and category “A” and this means that SNOMED-CT concept model allows multiple-inheritance. Today, the International Health Terminology Standards Development Organization (IHTSDO) (http://www.ihtsdo.org) is responsible for the development of SNOMED-CT, for quality issues and the distribution of the terminology.

Table 1
First-level categories of SNOMED-CT concepts

First-level category, abstract	Size (number of concepts)
“Clinical finding” (finding)	109,311
“Special concept” (Special concept)	67,342
“Procedure” (procedure)	53,854
“body structure” (body structure)	31,837
“organism” (organism)	27,952
“Substance” (Substance)	23,456
“Pharmaceutical/biologic product” (product)	19,084
“Qualifier value” (Qualifier value)	8904
“Event” (event)	8447
“Observable entity” (observable entity)	7834
“Social context” (social concept)	5252
“Situation with explicit context” (situation)	4912
“physical object” (physical object)	4515
“Environment or geographical location” (environment/location)	1741
“Linkage concept” (linkage concept)	1136
“Staging and scales” (Staging scale)	1113
“Specimen” (specimen)	1055
“Record artifact” (record artifact)	202
“Physical force” (physical force)	172
Total concepts	378,111

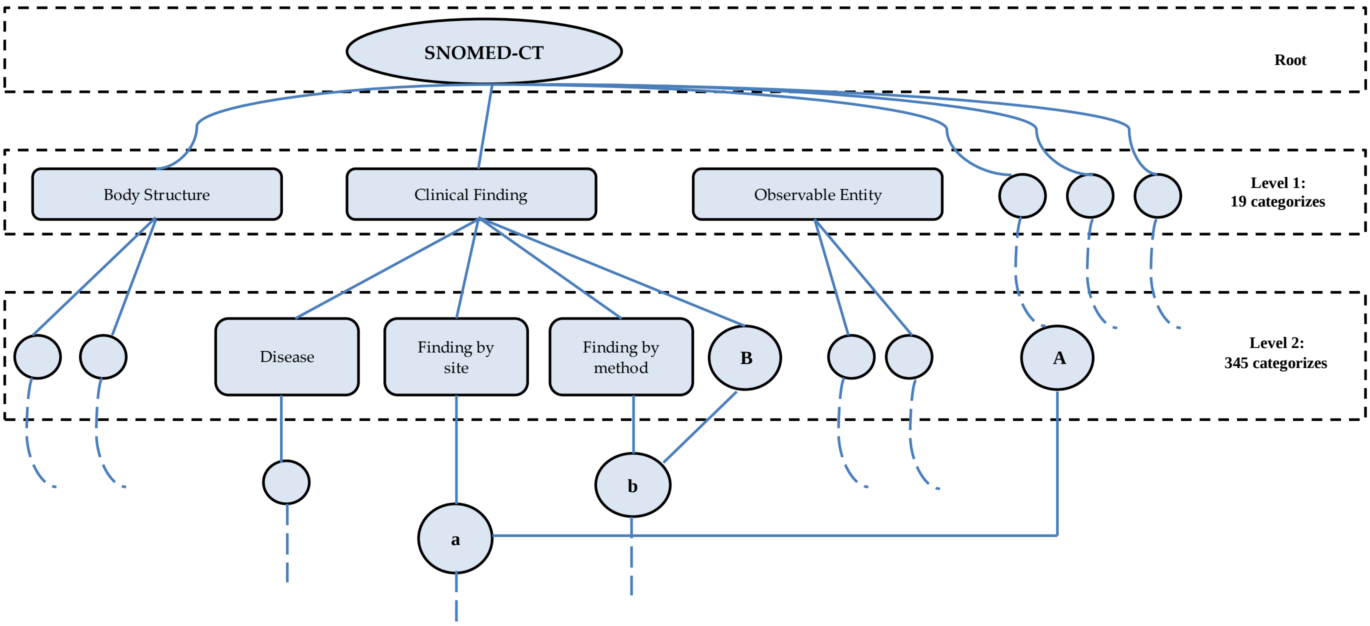


Fig. 1. Hierarchical representation of SNOMED-CT concepts

An International Journal JSCDSS E-ISSN: 2289-8603

2.1.2 UMLS (Unified Medical Language System)

The Unified Medical Language System (UMLS) was developed by the US National Library of Medicine and is a set of files and software that brings together many health and biomedical vocabularies to enable interoperability between computer systems (El-Rab et al., 201; Merabti et al., 2010; Bodenreider, 2004). In details, UMLS consists of three knowledge sources and a set of software tools which can be applied to access these knowledge sources. Knowledge sources are the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. One such Metathesaurus is the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). Semantic network in UMLA consists of a set of broad subject categories to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. In UMLS, the SPECIALIST Lexicon includes terms with linguistic information where its applications are in the domain of biomedical and healthcare (Chen et al., 2009; Marquet et al., 2007). UMLS can be freely accessed for the research purpose, but a license is needed.

3. Semantic similarity measures

In this section, we classify the similarity measures into two broad categories: path-based and information content (IC)-based. The path-based similarity measures provide information about the co-location of the terms in a taxonomy (taxonomy refers to the particular classification). A taxonomy that would be suitable for semantic similarity measures applications can be derived from a knowledge source like an ontology. The IC-based measures use the taxonomy information, but also include additional information about the concept with respect to its relationship with the other concepts. There are two methods to calculate IC: corpus-based, which uses the probability of a concept occurring in a corpus, and intrinsic-based, which uses the informativeness of a concept, based on its placement within the taxonomy. The remainder of this section describes the various semantic similarity measures and how they are calculated.

3.1 Path-based measures

Rada et al. (1989) introduced the Conceptual Distance measure, which is the length of the shortest path (spath) between two concepts (c1 and c2) using broader-narrower relations. Caviedes and Cimino (2004) later evaluated this measure using the parent-child relation.

The path measure is a modification of this and is calculated as the reciprocal of the length of the shortest path, as defined in Eq. (1).

$$sim_{path} = \frac{1}{spath(c_1, c_2)} \tag{1}$$

Wu and Palmer (1994) extended this measure by incorporating the depth of the Least Common Subsumer (LCS). The LCS is the most specific ancestor that two concepts shares. In this measure, the similarity is twice the depth of the two concepts' LCS, divided by the product of the depths of the individual concepts, as defined in Eq. (2).

$$sim_{wup} = \frac{2 * depth(LCS(c_1, \mathrm{c}_2))}{depth(c_1) + \mathrm{depth}(\mathrm{c}_2)} \tag{2}$$

Leacock and Chodorow (1998) used the depth of the taxonomy and developed a new path-based measure. In this method, the similarity between two given concept (c1, c2) is the negative log of the shortest path (spath) between them, divided by twice the total depth of the taxonomy (D) as shown in Eq. (3).

$$sim_{kh} = -\log \frac{spath(c_1, \mathrm{c}_2)}{2 * D} \tag{3}$$

Nguyen and Al-Mubaid (2006) proposed a new path-base measure by incorporating both the depth of the taxonomy and LCS of two given concepts (c1, c2) in their measure. In this measure, the similarity is defined as the log of two plus the product of the shortest distance between the two concepts minus one, multiplied by the subtraction of depth of the concepts' LCS (d) from the depth of the taxonomy (D). This measure is shown in Eq. (4) and its range depends on the depth of the taxonomy.

$$sim_{nam} = \log(2 + (\min path(c_1, \mathrm{c}_2) - 1) * (D - d)) \tag{4}$$

3.2 Information Content (IC) measures

Resnik (1999) used IC of concept to be applied in similarity measure. In this measure, the similarity of two concepts (c1, c2) is defined as the IC of their LCS, as shown in Eq. (5).

$$sim_{res} = IC(LCS(c_1, \mathrm{c}_2)) = -\log(P(LCS(c_1, \mathrm{c}_2))) \tag{5}$$

Jiang and Conrath (1997) and Lin (1998) redefined Resnik (1999) similarity measure and incorporated the IC of individual concepts in that method.

Lin (1998) proposed a new IC-based measure and defined the similarity between two concepts (c1, c2) as: dividing twice the IC of the concepts' LCS by sum of the individual IC of each concept (see Eq. (6)).

$$sim_{lin} = \frac{2 * IC(LCS(c_1, \mathrm{c}_2))}{IC(\mathrm{c}_1) + \mathrm{IC}(\mathrm{c}_2)} \tag{6}$$

An International Journal JSCDSS
E-ISSN: 2289-8603

This measure is very similar to path-based measure, proposed by Wu and Palmer (1994); but using IC of a concept instead of using its depth.

Jiang and Conrath (1997) developed a new IC-based measure by defining the distance between two concepts (c1, c2) to be the sum of the individual IC of concepts minus twice the IC of the concepts' LCS (See Eq. (7)).

$$sim_{jcn} = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(LCS(c_1, c_2))} \quad (7)$$

3.3 Information content

The information content of a concept can be calculated using information derived from a corpus (corpus-based) or information derived from a taxonomy (intrinsic-based). In this section, we describe both techniques.

In corpus-based, the IC of a concept is defined as the negative log of the probability of a concept as defined in Eq. (8).

$$IC(\mathrm{c}) = -\log(\mathrm{P(c)}) \quad (8)$$

Then, the probability of concept c $(P(c^*))$ is calculated by summing the probability of the concept itself occurring in some text (P(c)), plus the probability of its descendants $(\sum_{descendant(c)} P(d))$, occurring in the same text as seen in Eq. (9).

$$P(\mathrm{c}^*) = P(\mathrm{c}) + \sum_{descendant(c)} P(d) \quad (9)$$

Where P(c), initial probability of a concept, is calculated through dividing *freq*(d) by N (see Eq. (10)). In the following equation, N and *freq*(d) indicate the total number of concepts in the corpus and the number of times a concept is seen in the corpus, respectively.

$$\begin{aligned} P(c) &= freq(c) / N \\ P(\mathrm{d}) &= \mathrm{freq(d)} / \mathrm{N} \end{aligned} \quad (10)$$

A sufficient coverage of a taxonomy to obtain accurate estimation is a main challenge in probability calculations of concepts. Hence, to overcome this issue, intrinsic IC calculation which is based on ontologies has been proposed by Sanchez et al. (2011). In this approach the IC of a concept is assessed by its informativeness according to concept location in the hierarchy, considering its ancestors (incoming) and descendants (outgoing) (see Eq. (11)).

$$IC(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{\mathrm{max\_leaves} + 1}\right) \quad (11)$$

where leaves are the number of descendants of concept c that are leaf nodes, subsumers are the number of concept c's ancestors and *max* leaves are the total number of leaf nodes in the chosen taxonomy.

4. Academic papers summarization

In this section, we summarized and classified the studies that applied semantic similarity measures in biomedical domain using SNOMED-CT.

4.1 Path-based measures

Batet et al. (2011) proposed a new path-based semantic similarity measure, capturing more semantic evidence than previous path-based methods, based on the exploitation of ontologies' taxonomical structure. In this study SNOMED CT (as an input ontology) was applied to evaluate the accuracy of their proposed measure and a standard benchmark was used to compare this measure against other approaches. The correlation between the results of the evaluated measures and the human experts' ratings showed that their proposal outperformed all previous path-based measures, avoiding at the same time some of their limitations.

Caviedes and Cimino (2004) proposed a conceptual matching method to assess the similarity between two given concepts according to the minimum number of parent links between the two concepts. The performance of the proposed measure was evaluated based on domain experts' judgments on three sets of concepts, exploiting from UMLS knowledge source (SNOMED-CT, MeSH and ICD9CM). They argued that "by identification of semantically similar concepts, conceptual matching enables reasoning in the absence of exact lexical matching". As stated by authors, conceptual matching can also be applied in terminology development and maintenance, decision support system development, machine learning and data mining research in other fields.

Martínez et al. (2013) argued that the nature of structured patient data, like EHRs require some anonymization procedure for privacy purposes before releasing them to third parties. In order to address this issue, privacy preserving methods (such as Statistical Disclosure Control (SDC) techniques) have been recently proposed. However, most of these methods focus on continuous-scale numerical data without consideration that part of data in EHRs that is expressed with non-numerical attributes. Therefore, SDC application to EHRs produces are far from optimal results. In this regards, authors proposed a general framework using a path-based semantic similarity method to enable the accurate application of SDC

An International Journal JSCDSS
E-ISSN: 2289-8603

techniques to non-numerical part of clinical data. In this study, Batet et al. (2011) path-based method was used to exploit SNOMED CT as a structured medical knowledge source, helping to aggregate and sort non-numerical terms. Accordingly, their proposed framework was employed to several well-known SDC techniques to evaluate its performance by using a real clinical dataset containing non-numerical attributes. Results showed that the proposed approach has a high potential to produce anonymized datasets which better preserves the utility of EHRs.

Al-Mubaid and Nguyen (2006) proposed a new distance semantic similarity measure for the biomedical domain within the framework of UMLS (MeSH and SNOMED-CT). The proposed technique not only was able to take the path length feature into account, but it also used the depth of concept nodes to highly improve performance. The main contribution of this study was proposing a path-based measure with new attributes (common specificity and local granularity) that incorporated non-linearly feature in the proposed method. The experimental results indicated the efficiency of the proposed method and its high correlation with human judgments.

Al-Mubaid and Nguyen (2009) developed a path-based semantic similarity method to measure similarity between concepts using multiple ontologies (MeSH and SNOMED CT) in order to address the issue of flawed exploiting concepts from a single ontology. The proposed measure was based on three principal features, including: "cross-modified path length between two concepts, a new feature of common specificity of concepts (LCS) in the ontology, and local granularity of ontology clusters". The experimental results validated the efficiency of the proposed technique in single and multiple ontologies and its high correlation with human judgments.

Batet et al. (2013) stated that most previous works on semantic similarity supported only a unique input ontology, while knowledge is dispersed through several partial and/or overlapping ontologies in many domains. Hence, they proposed an ontological structure-based method to enable similarity estimation across multiple ontologies (SNOMED-CT, MeSH and WordNet). As stated by authors, the proposed method allows estimating the similarity between two terms when one is missing in a certain ontology, but it might find in another ontology. This can be done by discovering common concepts between two given terms that could act as bridges between different ontologies. In addition, in case of overlapping knowledge, which means several ontologies covering the same pair of terms, this approach is able to improve the accuracy of similarity estimation between terms by selecting the most accurate similarity assessment from all assessments computed by ontologies. Results of evaluation showed that their method was able to improve the accuracy of similarity estimation in comparison to single input ontology approaches.

4.2 Information Content (IC)-based

SáNchez and Batet (2011) presented new semantic similarity measures expressed in terms of Information Content (IC) of concepts. In this study they redefined several well-known edge-based measures in addition to a number of similarity coefficients in terms of IC, obtaining new semantic similarity functions. Their new IC-based methods used the taxonomic structure of biomedical ontologies like SNOMED CT to compute IC. They believed that the proposed approach which is based on ontology and intrinsic IC computation is capable to overcome the limitation of a suitable corpora availability. This approach assumes that the taxonomic structure of ontologies is organized in a meaningful way.

Batet et al. (2014) proposed a new IC-based approach to assess the similarity of two given concepts, spreading throughout several ontologies. They believed that the applicability of IC-based measures is hampered, if they solely deal with a single input ontology to compute similarity of concepts. This limitation can be overcome by multi-scenario ontologies; especially in the domain of biomedical that several knowledge sources are available. Therefore, they proposed that IC-based method to enable an accurate IC-based similarity assessment using multiple ontologies. The structure of their approach was based on the Information Theory, looking for the available subsume pair, that can act as the best MICA (Most Informative Common Ancestor) for the compared concepts, across multiple ontologies (SNOMED-CT and MeSH).

Fan & Friedman (2007) developed a corpus-based method based on distributional properties of terms to facilitate the semantically classification of ontological concepts for Natural Language Processing (NLP) applications. In this study, authors particularly focused on reclassifying UMLS (Level1+SNOMED-CT) concepts into broader semantic classes in order to develop a more classified UMLS structure for the needs of NLP applications. Apparently, the proposed method could also be used to improve the ontology itself and the performance of the systems depending on it. They argued that such an approach differs significantly from the classical methods that experts classify ontological concepts manually. Results also acknowledged that the proposed approach can recommend high level semantic classification, suitable for use in natural language processing.

Saruladha et al. (2011) proposed an IC-based semantic similarity measure and corpus independent based on Tversky model to assess similarity among cross ontological concepts. In this study, they also refined Resnik (1999) and Lin (1998) measures to compute cross ontological semantic similarity. The three proposed methods were evaluated with two biomedical ontologies, SNOMED-CT and MeSH within UMLS Framework, and tested with human judgments to assess their performance. Results showed their high efficiency in multi-scenario ontologies, as they achieved high correlation rates with experts' ranking. It was claimed that the proposed methods could be applied

An International Journal
JSCDSS
E-ISSN: 2289-8603

for ontology mapping, ontology alignment and information retrieval.

4.3 Hybrid semantic similarity measures

Gøeg et al. (2015) used Lin (1998) and Sokal and Sneath (1963) IC-based measures, with two aggregation techniques (All-pair AVG and Best-pair AVG), resulting in a total of four methods (Lin/AllAVG, Lin/BestAVG, SoSn/AllAVG and SoSn/BestAVG) in order to harmonize and standardize clinical models. In this study, SNOMED-CT was chosen because of its coverage and flexibility compared to other terminologies and to obtain an intrinsic-similarity estimation. It was claimed that the study can support hospitals by proposing them guidance in order to change or create templates for the purpose of harmonization.

García et al. (2012) applied path-based techniques in combination with other approaches to bind OpenEHR archetype terms to an external terminology, SNOMED CT. They employed path-based methods to validate the bindings, resulting from lexical techniques, and to resolve ambiguous binding conflicts.

Mabotuwana et al. (2013) presented a new semantic vector based on the semantic distance between sets of concepts (instead of individual concepts) to determine similarity between two given documents using SNOMED CT as a reference ontology. In this approach, the notion of edge-based semantic similarity (taking advantage of the IS-A relations) was used in vector space model to overcome the limitations of Direct Concept Matching (DCM). DCM is matching the exact same concepts in the two compared documents. They tested and evaluated the proposed approach in classification of radiology reports into anatomy and procedure-based groups. The evaluation showed that the proposed semantic approach increases the similarity of documents describing the same anatomies. This led to improving classification accuracy of documents, compared to a non-semantic approach.

Pivovarov and Elhadad (2012) proposed a comprehensive method, which computes a similarity score for a concept pair by combining data-driven and ontology-driven knowledge. In this paper, they examined the problem of concept aggregation in the context of a clinical data-mining task. For example, concepts such as ‘‘obese’’ and ‘‘morbidly obese’’ can be merged when studying Huntington’s disease, but should remain separate when investigating predictors for heart attack. Regarding this, a homogenous corpus of notes (notes about patients who share at least one clinical problem) were preprocessed to extract related concepts. Then, In order to prune out the extracted concepts and achieve a homogeneous set of them for aggregation, a three-way filter was employed. Next, a context-based similarity method estimated the similarity between all pairs of concepts. Finally, the top-k pairs with the highest context-based similarity were reordered using the two knowledge-based similarity measures. In this study, the proposed method was applied on concepts from SNOMEDCT and a corpus of patients’ clinical notes, containing chronic kidney disease. The authors claimed that their work fits well within the field of clinical informatics to enrich the analysis of unstructured data located in EHRs.

Garla and Brandt (2012b) developed a novel context-sensitive semantic similarity measure by combining feature ranking and semantic similarity methods to support clinical document classification. The critical steps in text classification consist of features identification relevant to the classification task, and representation of text to enable discrimination between documents of different classes. Hence, in this study, a new feature ranking method was presented to utilize the knowledge encoded in the taxonomy of UMLS (SNOMED-CT, ICD-9, and RXNORM). The Lin (1998) measure is also employed to compute concepts similarity, as it showed a high correlation with expert judgments in empirical evaluations. They argued that our “context dependent” semantic similarity measure tailors the “perception” of similarity to a specific classification task which improves the performance of machine learning techniques in clinical text classification.

Steichen et al. (2006) built an ontology of morphological abnormalities in breast pathology to assist inter-observer consensus. First, the concepts of this ontology extracted from medical sources, such as medical reports and ontologies. SNOMED CT, GALEN, and GeneOntolog were the selected ontologies that contain pathological concepts. Next, the extracted concepts were organized in a taxonomic hierarchy and linked by the IS-A relation based on diagnostic meaning. After creating the ontology, a validation process was performed to examine the quality of the ontology. In this stage, a set of semantic similarity measures including, position-based and IC-based were applied between concepts and their results were evaluated according to experts’ judgments.

Table 2 shows the studies of semantic similarity measures that used SNOMED-CT as an input ontology.

5. Reference standards in biomedical domain

By evolving different semantic similarity measures in biomedical domains, efforts have been done to evaluate these measures. Semantic similarity measures are normally evaluated by means of standard benchmarks of word pairs whose similarity has been assessed by a group of human experts. In this process, the correlation of similarity values (obtained by SSM) and human similarity estimation is calculated. The correlation ranges from 0 to 1 and if the correlation is near to 1, it indicates that the measure properly approximates the judgments of human, which is precisely the goal.

There are several benchmarks in biomedical domain which have been used to evaluate semantic similarity measures performance in the recent years. One of the well-known benchmark in the biomedical domain is the benchmark created by Pedersen et al. (2007). It consists of 30 pairs of SNOMED CT concepts whose similarities were

An International Journal JSCDSS
E-ISSN: 2289-8603

assessed by experts of the Mayo Clinic. In this benchmark, a total of 12 experts, including three physicians and nine medical coders, assessed each word pair similarity. After a normalization process, the average similarity values between each word pair was provided in a scale, ranging from 1 (non-similar) to 4 (identical).

Table 2
Semantic similarity measures that used SNOMED-CT as an input ontology

Author	Approach	About the proposed measure
Batet et al. (2011)	Path-based	The proposed measure used a broad taxonomic knowledge.
Caviedes and Cimino (2004)	Path-based	The measure was relevant for ontology maintenance and development, as well as for machine learning and data mining research in biomedical informatics.
Martínez et al. (2013)	Path-based	The measure was used to improve Statistical Disclosure Control (SDC) methods for enhancing Anonymization procedure of EHRs.
Al-Mubaid and Nguyen (2006)	Path-based	The proposed method used depth of concepts nodes to improve SSM performance.
Al-Mubaid and Nguyen (2009)	Path-based	The method used the depth and length of the path between concepts to measure the similarity in single ontology and across multiple ontologies.
Batet et al. (2013)	Path-based	The proposed measure supports similarity estimation across multiple ontologies.
SáNchez and Batet (2011)	IC-based	The method enables medical data classification such as clinical records. It also helps the integration of heterogeneous clinical data (like clinical records expressed in different formats).
Batet et al. (2014)	IC-based	The proposed method assesses the similarity of concepts spread throughout several ontologies to deal with cross domain data.
Fan & Friedman (2007)	IC-based	The measure can assist reclassifying UMLS concept and support maintaining and developing ontologies.
Saruladha et al. (2011)	IC-based	This proposed method is based on Tversky's SSM model and relevant for ontology mapping and ontology alignment.
Gøeg et al. (2015)	Hybrid	The measure was used to harmonize and standardize clinical models. It can propose hospitals a guidance to create templates for the purpose of harmonization.
García et al. (2012)	Hybrid	The proposed measure has the potential to resolve ambiguous archetypes binding conflicts.
Mabotuwana et al. (2013)	Hybrid	The measure supports data classification, and in this study, it was used to support classification of radiology reports.
Pivovarov and Elhadad (2012)	Hybrid	The measure used to solve the problem of concept aggregation in clinical data-mining task. It enabled the analysis of unstructured data located in EHRs.
Garla and Brandt (2012b)	Hybrid	The method improves the performance of machine learning techniques, so as to support classification.
Steichen et al. (2006)	Hybrid	The proposed measure supports maintaining and developing ontologies.

Pakhomov et al. (2011) and Pakhomov et al. (2010) developed larger benchmarks Mayo and UMN respectively for evaluating semantic similarity and relatedness measures using UMLS medical concept pairs. In 'Mayo' benchmark, the same nine medical coders and three physicians, who supplied rating for the Pedersen et al. (2007) benchmark, assessed the semantic relatedness of 101 UMLS word pairs on an ordinal scale. In the 'UMN' benchmark, eight medical residents ranked a set of 587 and 566 UMLS concept pairs on a continuous scale for relatedness and similarity respectively. Hliaoutakis (2005) also developed a benchmark in biomedical domain, containing a set of 36 medical terms extracted from the MeSH ontology. In this benchmark the similarity between each word pair was ranked by eight medical experts, ranging from 0 (non-similar) to 1 (identical).

6. Evaluation of semantic similarity measures

In this section, we present studies that evaluated semantic similarity measures according to different benchmarks. Academic papers presented in this review have mostly employed the above mentioned benchmarks, using single ontology (SNOMED-CT) and in some cases, multiple ontologies (such as SNOMED-CT and MeSH).

Batet et al. (2011) used the Pedersen et al. (2007) benchmark and SNOMED CT ontology to evaluate the accuracy of their proposed path-based measure. They also presented an objective comparison between their proposed measures and other measures in the biomedical domain. Based on results and considering the correlation values between human experts in Pedersen et al. (2007) benchmark (0.68 for physicians and 0.78 for coders), the proposed measures performed comparatively well, obtaining a high correlation with human judgments.

McInnes and Pedersen (2015) evaluated the recent semantic similarity and relatedness measures to identify a pair of measures so as to improve the accuracy of similarity assessment between two terms. In this study, SNOMED CT taxonomy was applied as an input ontology for the similarity measures and the entire UMLS (Level 1+SNOMED CT) for the relatedness measures. They evaluated the measures based on 3 standard benchmarks, including Pedersen et al. (2007) and Pakhomov et al. (2010, 2011) benchmarks (Mayo and UMN). Results showed that combining relatedness and similarity measures more closely correlates with human judgments; especially using Lesk (1986) as a relatedness measure and Jiang and Conrath (1997), obtained the highest overall correlation with reference standards.

Gøeg et al. (2015) used IC-based measures of Lin (1998) and Sokal and Sneath (1963), with two aggregation techniques (All-pair AVG and Best-pair AVG), resulting in a total of four methods (Lin/AllAVG, Lin/BestAVG, SoSn/AllAVG and SoSn/BestAVG). Evaluation results showed that the two IC-based similarity measures with BestAVG aggregation technique have the highest potential of clustering similar templates based on generated dendrograms. However, no difference was seen in choice of Lin (1998) and Sokal and Sneath (1963) IC-based measures.

In the study by SáNchez and Batet (2011), an objective comparison between several IC-based and non IC-based measures proposed, using Pedersen et al. (2007) benchmark and SNOMED CT as a knowledge source. They argued, "the fact that Pedersen et al. (2007) benchmark and SNOMED CT have become almost de facto evaluation standards in recent works, allows a fair evaluation and a clear comparison". Results showed that redefinition of non IC-based measures in terms of IC with an intrinsic estimation (from SNOMED-CT) led to a noticeable performance improvement. The best accuracy has occurred in re-formulated of Sokal and Sneath (1963) method from set-based to IC-based with intrinsic calculation.

Sánchez and Batet (2013) argued that IC-based approaches which compute IC of concepts in intrinsic manner have promising results in computing similarity between terms compared to other paradigms used by related works.

Sánchez and Batet (2013) claimed that intrinsic IC-based approaches have shown a great performance in assessing the similarity of two terms; however, these approaches are largely hampered by the coverage offered by the single input ontology. In this regards, the above limitation could be overcome by computing IC of concepts from multiple ontologies. Therefore, they applied well-known IC-based similarity measures such as Resnik (1995), Lin (1998), and Jiang and Conrath (1997) by considering multiple ontologies (SNOMED CT and MeSH) in an integrated way. Next, Pedersen et al. (2007) benchmark was used in order to provide an objective evaluation between these measures and related works. Results showed that, first of all intrinsic IC-based measures (with single input ontology) obtained higher correlation values with human judgment than edge-counting measures and measures based on corpora. Moreover, the exploitation of several complementary and/or overlapping ontologies during the similarity assessment improved significantly the accuracy of the IC-based measures compared to the single input ontology.

In the study by Batet et al. (2014), a new IC-based measure was proposed and an empirical evaluation used based on well-established benchmarks in biomedical ontologies. In order to do so, several similarity measures were compared against the proposed method based on the correlations with Pedersen et al. (2007) and Pakhomov et al. (2010, 2011) benchmarks and using SNOMED-CT and MeSH as ontologies. Results of evaluation showed the higher accuracy and noticeable performance of their approach in relation to related works.

Steichen et al. (2006) conducted an evaluation study to compare the three similarity measures based on experts' judgment. All three measures were well matched with the experts' judgment and none of them was better than the other. The Leacock and Chodorow (1998) path-based measure using both taxonomic and non-taxonomic links performed as well as Lin (1998) and Jiang and Conrath (1997).

McInnes and Pedersen (2013) developed a Word Sense Disambiguation (WSD) method that was able to disambiguate terms in biomedical text, using semantic similarity and relatedness measures which extracted information from MeSH and entire UMLS (Level1+SNOMED-CT). Regarding this, path-based, corpus-based, intrinsic IC-based measures, and relatedness measures were compared based on the Pedersen et al. (2007) and Hliaoutakis (2005) benchmarks to find the quality and efficacy of them on WSD. The overall results showed that IC-based measures (especially Lin (1998)) derived from either a corpus or a taxonomy, obtained higher disambiguation accuracy than the other measures.

In the study by Al-Mubaid and Nguyen (2006), the proposed measure (known as "Sem") was evaluated based on Pedersen et al. (2007) benchmark and compared with five ontology-based similarity measures, including: Rada et al. (1989), Wu & Palmer (1994), Leacock & Chodorow (1998), Li et al. (2003) and Choi and Kim (2003) using MeSH and SNOMED-CT as input ontologies. Results showed that the proposed method has achieved the best overall correlation score with human ratings and so proved its efficiency. In addition, using MeSH ontology rather than SNOMED-CT produced better semantic correlations with human ratings in all of six tested measures.

An International Journal JSCDSS
E-ISSN: 2289-8603

Al-Mubaid and Nguyen (2009) conducted an experimental evaluation to evaluate their proposed measure ("SemDist") relative to human similarity scores such as Pedersen et al. (2007) and Hliaoutakis (2005) benchmarks. The proposed measure was also compared with other similarity measures such as Leacock and Chodorow (1998), Wu and Palmer (1994) and Rada et al. (1989) (Path length) methods, using MeSH and SNOMED CT. The experimental results confirmed the efficiency of the proposed measure and its best correlations with human scores within both benchmarks. Results also showed that MeSH terminology produces better semantic similarity correlations with human ratings than SNOMED-CT in all of the tested measures.

Batet et al. (2013) proposed a new path-based method and evaluated it based on standard benchmarks of Hliaoutakis (2005) and Pedersen et al. (2007), using several ontologies (SNOMED-CT, MeSH and WordNet). Moreover, an objective comparison in multiple-ontological scenarios was conducted to compare the proposed method with "SemDist", proposed by Al-Mubaid and Nguyen (2009). The experimental results validated the efficiency of the proposed technique and its higher correlation results with human scores in relation to SemDist. The argued that on the contrary to SemDist who relied on the path to compute similarities, our measure considers all taxonomical ancestors. Hence, these theoretical advantages have been reflected in improved similarity accuracy when compared with "SemDist".

In a comprehensive study conducted by Garla and Brandt (2012a), the performance of a number of semantic similarity measures were evaluated within the UMLS framework (SNOMED-CT and MeSH) using three different benchmarks; including, Mayo, UMN (Pakhomov et al. (2010, 2011) benchmarks) and Pedersen et al. (2007). Results showed that among different types of semantic similarity measures, ontological-based approaches (path-based and intrinsic IC-based) outperformed the corpus IC-based techniques. Among ontological-based approach, intrinsic IC-based significantly outperformed path-based measures.

Table 3 presents these studies that evaluated semantic similarity measures according to different benchmarks.

Table 3
The list of studies that evaluated semantic similarity measures based on biomedical benchmarks

Study	Best Measure	Type of Measure	Benchmark
Batet et al. (2011)	Their proposed measure	Path-based	Pedersen et al. (2007)
McInnes and Pedersen (2015)	Lesk(1986) with Jiang and Conrath (1997) methods Combination	Jiang and Conrath (1997) method is an IC-based measure; Lesk (1986) method is a relatedness measure.	Pedersen et al. (2007) and Pakhomov et al. (2010, 2011)
Gøeg et al. (2015)	No difference in choice of Lin (1998) and Sokal and Sneath (1963) IC-based measures	…	Dendrograms (to check the classification performance assisted by SSM)
SáNchez and Batet (2011)	"Sokal and Sneath" Reformulation	IC-based measure	Pedersen et al. (2007)
Sánchez and Batet (2013)	Resnik (1995), Lin (1998), Jiang and Conrath (1997)	IC-based measure and using multiple ontologies	Pedersen et al. (2007)
Batet et al. (2014)	Their proposed measure	IC-based	Pedersen et al. (2007); Pakhomov et al. (2010, 2011)
Steichen et al. (2006)	No compared measure performed better (all compared measures performed almost same)	….	Local experts
McInnes and Pedersen (2013)	Lin (1998)	IC-based measure and using multiple ontologies	Pedersen et al. (2007) and Hliaoutakis (2005)
Al-Mubaid and Nguyen (2006)	Their proposed measure ("Sem")	Path-based	Pedersen et al. (2007)
Al-Mubaid and Nguyen (2009)	Their proposed measure ("SemDist")	Path-based measure and using multiple ontologies	Pedersen et al. (2007) and Hliaoutakis (2005)
Batet et al. (2013)	Their proposed measure	Path-based measure and using multiple ontologies	Pedersen et al. (2007) and Hliaoutakis (2005)
Garla and Brandt (2012a)	Intrinsic IC-based measures	IC-based measures and using multiple ontologies	Pedersen et al. (2007); Pakhomov et al. (2010, 2011)

An International Journal JSCDSS E-ISSN: 2289-8603

7. Discussion

The most significant point that we can draw from evaluation studies is that no clear best measure was found for assessing the similarity of concepts. In different situation, measures performed differently. It means that a measure was performed poorly for a given task, while suited well in another. As an example, “Lin” IC-based measure was found by McInnes and Pedersen (2013) to be the most accurate measure in assessing the degree of similarity between two terms, while Batet et al. (2014) found it to be fairly week when evaluated it against their approach.

In addition, from this study, it was found that there is no clear best strategy for evaluating semantic similarity measures. For example, Pedersen et al.’s (2007) benchmark used the SNOMED-CT 2004 version; however, the SSMs proposed by the recent studies have used the updated version of SNOMED-CT to assess the similarity between concepts. Therefore, these comparisons may be flawed. One important step to tackle this issue would be the development of a gold standard benchmark that allows the effective evaluation of semantic similarity measures.

On the other hand, ontological-based semantic similarity measures are dependent on the quality and completeness of an ontology. Meanwhile, biomedical ontologies (such as SNOMED-CT), have several irregularities such as variable edge length, variable depth, and variable node density (Pesquita et al., 2009). Therefore, these methods should take these irregularities into account, since their accuracy can be affected by these issues.

8. Conclusion

Over the last decade, ontologies have become an increasingly important component of biomedical studies. It is due to the fact that they provide researchers with common terminologies to present information in a structured way, and so to be shared and reused easily by humans and computers. The main benefit of ontologies is that the hierarchy of concepts can be objectively compared through semantic similarity measures.

This study offered a broad overview regarding ontologies and semantic similarity measures. Besides, it examined the current situation of SSM in biomedical domain based on a review of related works. We found that IC-based measures were typically more reliable, achieving higher correlation with experts’ judgment, compared to path-based methods.

In the biomedical domain no comprehensive study has been conducted in the recent years to evaluate states-of-the-art measures with related works. Therefore, in the future, we aim to conduct a comprehensive study to evaluate all the proposed semantic similarity methods in biomedical domain, by using several ontologies and standard benchmarks; so as to gain a clearer insight about the performance of each method.

References

Al-Mubaid, H., & Nguyen, H. (2006, August). A cluster-based approach for semantic similarity in the biomedical domain. In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE (pp. 2713-2717). IEEE.

Al-Mubaid, H., & Nguyen, H. (2009). Measuring semantic similarity between biomedical concepts within multiple ontologies. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 39(4), 389-398.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., & Ranwez, V. (2014). An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Information Sciences, 283, 197-210.

Batet, M., Sánchez, D., & Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. Journal of biomedical informatics, 44(1), 118-125.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2013). Semantic similarity estimation from multiple ontologies. Applied intelligence, 38(1), 29-44.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far.Semantic Services, Interoperability and Web Applications: Emerging Concepts, 205-227.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), D267-D270.

Campbell, J. R., Brear, H., Scichilone, R., White, S., Giannangelo, K., Carlsen, B., ... & Fung, K. W. (2013). Semantic interoperation and electronic health records: context sensitive mapping from SNOMED CT to ICD-10. In MedInfo (pp. 603-607).

Caviedes, J. E., & Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the UMLS. Journal of biomedical informatics,37(2), 77-85.

Chakraborty, U. K., Gurung, R., & Roy, S. (2014, December). Semantic Similarity Based Approach for Automatic Evaluation of Free Text Answers Using Link Grammar. In Technology for Education (T4E), 2014 IEEE Sixth International Conference on (pp. 218-221). IEEE.

Chang, J. Y., & Lee, K. M. (2015). Large margin learning of hierarchical semantic similarity for image classification. Computer Vision and Image Understanding, 132, 3-11.

Chaves-González, J. M., & MartíNez-Gil, J. (2013). Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowledge-Based Systems, 37, 62-69.

Chen, Y., Gu, H. H., Perl, Y., & Geller, J. (2009). Structural group-based auditing of missing hierarchical relationships in UMLS. Journal of biomedical informatics, 42(3), 452-467.

Choi, I., & Kim, M. (2003, July). Topic distillation using hierarchy concept tree. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 371-372). ACM.

Cross, V., Yu, X., & Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. International Journal of Approximate Reasoning, 54(7), 861-875.

Duarte, J., Castro, S., Santos, M., Abelha, A., & Machado, J. (2014). Improving Quality of Electronic Health Records with SNOMED. Procedia Technology, 16, 1342-1350.

El-Rab, W. G., Zaïane, O. R., & El-Hajj, M. (2013). Analyzing the Impact of UMLS Relations on Word-sense Disambiguation Accuracy. Procedia Computer Science, 21, 295-301.

An International Journal JSCDSS E-ISSN: 2289-8603

Fan, J. W., & Friedman, C. (2007). Semantic classification of biomedical concepts using distributional similarity. Journal of the American Medical Informatics Association, 14(4), 467-477.

García, M. M., Allones, J. L. I., Hernández, D. M., & Iglesias, M. J. T. (2012). Semantic similarity-based alignment between clinical archetypes and SNOMED CT: an application to observations. International journal of medical informatics, 81(8), 566-578.

Garla, V. N., & Brandt, C. (2012a). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC bioinformatics, 13(1), 261.

Garla, V. N., & Brandt, C. (2012b). Ontology-guided feature engineering for clinical text classification. Journal of biomedical informatics, 45(5), 992-998.

Gøeg, K. R., Cornet, R., & Andersen, S. K. (2015). Clustering clinical models from local electronic health records based on semantic similarity. Journal of biomedical informatics, 54, 294-304.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), 496.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. International journal of human-computer studies, 43(5), 907-928.

Guzzi, P. H., Mina, M., Guerra, C., & Cannataro, M. (2012). Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in bioinformatics, 13(5), 569-585.

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.

Iosif, E., & Potamianos, A. (2015). Similarity computation using semantic networks created from web-harvested data. Natural Language Engineering,21(01), 49-79.

Janowicz, K., Raubal, M., & Kuhn, W. (2015). The semantics of similarity in geographic information retrieval. Journal of Spatial Information Science, (2), 29-57.

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

Jiang, Y., Wang, X., & Zheng, H. T. (2014). A semantic similarity measure based on information distance for ontology alignment. Information Sciences,278, 76-87.

Lamurias, A., Grego, T., & Couto, F. M. (2013, October). Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In BioCreative Challenge Evaluation Workshop (Vol. 2, p. 75).

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. Journal of the American Medical Informatics Association, 21(e1), e11-e19.

Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. InProceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM.

Li, Y., Bandar, Z., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 15(4), 871-882.

Liao, H., Xu, Z., & Zeng, X. J. (2014). Distance and similarity measures for hesitant fuzzy linguistic term sets and their application in multi-criteria decision making. Information Sciences, 271, 125-142.

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

Mabotuwana, T., Lee, M. C., & Cohen-Solal, E. V. (2013). An ontology-based similarity measure for biomedical data–Application to radiology reports. Journal of biomedical informatics, 46(5), 857-868.

Marquet, G., Mosser, J., & Burgun, A. (2007). A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: the case of OBO disease ontologies. international journal of medical informatics, 76, S353-S361.

Martínez, S., Sánchez, D., & Valls, A. (2013). A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics, 46(2), 294-303.

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics, 46(6), 1116-1124.

McInnes, B. T., & Pedersen, T. (2015). Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. Journal of biomedical informatics, 54, 329-336.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

Merabti, T., Joubert, M., Lecroq, T., Rath, A., & Darmoni, S. J. (2010). Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH. Irbm, 31(4), 221-225.

Miller, T., Biemann, C., Zesch, T., & Gurevych, I. (2012). Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. In COLING (pp. 1781-1796).

Neches, R., Fikes, R. E., Finin, T., Gruber, T., Patil, R., Senator, T., & Swartout, W. R. (1991). Enabling technology for knowledge sharing. AI magazine, 12(3), 36.

Nguyen, H., & Al-Mubaid, H. (2006, May). New ontology-based semantic similarity measure for the biomedical domain. In Granular Computing, 2006 IEEE International Conference on (pp. 623-628). IEEE.

Pakhomov, S. V., Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., & Chute, C. G. (2011). Towards a framework for developing semantic relatedness reference standards. Journal of biomedical informatics, 44(2), 251-265.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., & Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. In AMIA annual symposium proceedings (Vol. 2010, p. 572). American Medical Informatics Association.

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

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. PLoS Comput Biol, 5(7), e1000443.

Pivovarov, R., & Elhadad, N. (2012). A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. Journal of biomedical informatics, 45(3), 471-481.

An International Journal
JSCDSS
E-ISSN: 2289-8603

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 19(1), 17-30.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res.(JAIR), 11, 95-130.

Sahni, L., Sehgal, A., Kochar, S., Ahmad, F., & Ahmad, T. (2014, December). A Novel Approach to Find Semantic Similarity Measure between Words. InComputational and Business Intelligence (ISCBI), 2014 2nd International Symposium on (pp. 89-92). IEEE.

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. (2013). A semantic similarity method based on information content exploiting multiple ontologies. Expert Systems with Applications, 40(4), 1393-1399.

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

Saruladha, K., Aghila, G., & Bhuvaneswary, A. (2011, June). COSS: Cross Ontology Semantic Similarity measure—An information content based approach. In Recent Trends in Information Technology (ICRTIT), 2011 International Conference on (pp. 485-490). IEEE.

Schulz, S., & Martínez-Costa, C. (2013). How ontologies can improve semantic interoperability in health care. In Process Support and Knowledge Representation in Health Care (pp. 1-10). Springer International Publishing.

Schulz, S., Rodrigues, J. M., Rector, A., Spackman, K., Campbell, J., Üstün, B., ... & Persson, K. B. (2014). What's in a Class? Lessons Learnt from the ICD–SNOMED CT Harmonisation. Studies in health technology and informatics, 205, 1038.

Sicilia, M. A. (Ed.). (2014). Handbook of metadata, semantics and ontologies. World Scientific.

Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy.Principles of numerical taxonomy.

Song, X., Li, L., Srimani, P. K., Yu, P. S., & Wang, J. Z. (2014). Measure the semantic similarity of go terms using aggregate information content.IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),11(3), 468-476.

Steichen, O., Daniel-Le Bozec, C., Thieu, M., Zapletal, E., & Jaulent, M. C. (2006). Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus. Computers in Biology and Medicine, 36(7), 768-788.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. Data & knowledge engineering, 25(1), 161-197.

Thomasson, A. L. (2014). Ontology made easy. Oxford University Press.

Uddin, M. N., Duong, T. H., Nguyen, N. T., Qi, X. M., & Jo, G. S. (2013). Semantic similarity measures for enhancing information retrieval in folksonomies. Expert Systems with Applications, 40(5), 1645-1653.

Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. InProceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

Zadeh, P. D. H., & Reformat, M. Z. (2013). Assessment of semantic similarity of concepts defined in ontology. Information Sciences, 250, 21-39.

Zaid, N. M., & Lau, S. K. (2014). Emerging of Academic Information Search System with Ontology-Based Approach. Procedia-Social and Behavioral Sciences, 116, 132-138.

An International Journal
JSCDSS
E-ISSN: 2289-8603