Journal of Biomedical Informatics xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain

Sébastien Harispe [a,*], David Sánchez [b], Sylvie Ranwez [a], Stefan Janaqi [a], Jacky Montmain [a]

[a] LGI2P/EMA Research Centre, Site de Nîmes, Parc scientifique G. Besse, 30035 Nîmes cedex 1, France
[b] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 16 July 2013
Accepted 9 November 2013
Available online xxxx

Keywords:
Ontologies
Semantic similarity measures
Unifying framework
SNOMED-CT
Biomedical ontologies

ABSTRACT

Ontologies are widely adopted in the biomedical domain to characterize various resources (e.g. diseases, drugs, scientific publications) with non-ambiguous meanings. By exploiting the structured knowledge that ontologies provide, a plethora of ad hoc and domain-specific semantic similarity measures have been defined over the last years. Nevertheless, some critical questions remain: which measure should be defined/chosen for a concrete application? Are some of the, a priori different, measures indeed equivalent? In order to bring some light to these questions, we perform an in-depth analysis of existing ontology-based measures to identify the core elements of semantic similarity assessment. As a result, this paper presents a unifying framework that aims to improve the understanding of semantic measures, to highlight their equivalences and to propose bridges between their theoretical bases. By demonstrating that groups of measures are just particular instantiations of parameterized functions, we unify a large number of state-of-the-art semantic similarity measures through common expressions. The application of the proposed framework and its practical usefulness is underlined by an empirical analysis of hundreds of semantic measures in a biomedical context.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Over the last decade, considerable efforts have been made to standardize our understanding of various fields by means of ontologies, i.e. formal and explicit specifications of shared conceptualizations [1]. Ontologies enable modelling domains through sets of concepts and semantic relationships established between them. Due to the importance of knowledge representation and terminology in biology and medicine, the biomedical domain has been very prone to the definition of structured thesauri or ontologies (e.g. UMLS, SNOMED-CT, MeSH). They enable characterizing medical resources such as clinical records, diseases, genes, or even scientific articles, through unambiguous conceptualizations. To take advantage of this valuable knowledge for information retrieval and knowledge discovery, *semantic similarity measures* are used to estimate the similarity of concepts defined in ontologies and, hence, to assess the semantic proximity of the resources indexed by them.

Ontology-based semantic similarity measures compare how similar the meanings of concepts are according to the taxonomical evidences modelled in the ontology. They are used in a wide array of applications: to design information retrieval algorithms [2,3], to disambiguate texts [4,5], to suggest drug repositioning [6] and to cluster genes according to their molecular function [7], to cite a few. Semantic similarity measures are indeed critical components of many knowledge-based systems [6,8,9]. Moreover, they are nowadays receiving more attention due to the growing adoption of both Semantic Web and Linked Data paradigms [10].

A plethora of measures have been proposed over the last decades (see surveys [7,9,11]). Although some context-independent semantic similarity measures have been proposed [12–15], most measures were designed in an ad hoc manner and were expressed on the basis of domain-specific or application-oriented formalisms [8]. Therefore, most proposals related to those measures target a specific audience and fail to benefit other communities. In this way, a non-specialist can only interpret the large diversity of state-of-the-art proposals as an extensive list of measures. As a consequence, the selection of an appropriate measure for a specific usage context is a challenging task. Actually, no extensive studies enabled characterizing the large diversity of proposals, even though few seminal contributions focusing on theoretical aspects of ontology-based semantic similarity measures exist [8,16,17].

Despite the large number of contributions related to ontology-based semantic similarity measures, the understanding of their foundations is nowadays limited. For a designer/practitioner, some fundamental questions remain: Why does a measure work better than another one? How does one choose or design a measure? Is it possible to distinguish families of measures sharing specific

* Corresponding author.
E-mail address: sebastien.harispe@mines-ales.fr (S. Harispe).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.11.006

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

properties? How can one identify the most appropriate measures according to particular criteria?

To fill these gaps, this paper proposes an extensive study of ontology-based semantic similarity measures from which a unifying framework decomposing measures through a set of intuitive core elements is proposed.

1.1. Contributions and plan

The framework presented in this paper proposes to model, in a generic and flexible way, the core elements on which most measures available in the literature rely. Thus, particular semantic measures can be properly characterized and can directly be obtained as instantiations of the framework components. This brings new insights for the study of semantic measures:

- *Distinguishing the core elements on which measures rely*. The theoretical characterization of semantic measures helps to understand the different measure paradigms and the large diversity of expressions proposed in the state-of-the-art.
- *Unifying measures through parameterized measures*. Based on the characterization of the core elements of semantic measures, our framework enables the identification of commonalities, bridges and equivalences between exiting measures. Indeed, their design could be unified through abstract expressions, even if many of them are (i) of ad hoc nature, (ii) domain-specific or (iii) based on different theoretical principles. Expressing semantic similarity measures through parameterized expressions can therefore facilitate the detection of their common properties and the analysis of their behaviour in specific applications.
- *Selecting appropriate domain-specific measures*. Such a framework provides a systematic, theoretically-coherent and direct way to define or tune the semantic similarity assessment for particular application scenarios. Semantic similarity measures expressed through parameterized functions could therefore be used to optimize measure tuning in domain-specific applications.
- *Designing new families of semantic measures*. New measures can be easily defined due to the modularity provided by the framework. Their design can take into account (i) the elements that affect the semantic assessment the most (e.g. estimation of concept specificity) and (ii) the particularities of ontology/application to which it will be applied (e.g. the presence of multiple inheritances).
- *Identifying the crucial aspects of semantic similarity assessment*. Empirical studies could be used to highlight the core elements best impacting measures' accuracies. As a result, the framework can be used to guide research efforts towards the aspects that can improve measure performances.

Such an approach will not just benefit a single measure designed for a domain-specific application (which is, to date, the focus of most related works) but will rather result in improvements of a wide set of measures and applications.

The rest of the paper is organized as follows. Section 2 introduces the reader to ontology-based semantic similarity measures, distinguishing the various paradigms proposed for their design. In addition, this section reviews previous works regarding the unification of semantic measures. Section 3 describes the proposed framework from which state-of-the-art measures are unified, and from which new proposals can be derived. Section 4 illustrates the practical application of the framework in which semantic measures' behaviours are analysed in a biomedical scenario. Section 5 provides the conclusions as well as some lines of future work.

2. Ontology-based semantic similarity measures

This section reviews the various paradigms used for the definition of ontology-based semantic similarity measures (SSMs). Each paradigm is illustrated by a selection of proposals emphasizing the essence of the approach. We then introduce the reader to existing contributions related to the unification of SSMs.

2.1. Paradigms for semantic similarity estimation

SSMs aim at estimating the likeness of two concepts considering the taxonomical knowledge modelled in ontologies. We consider approaches measuring taxonomic distance/dissimilarity indistinctly; notice that the latter can be converted to similarities by means of a linear transformation. In this section, we present state-of-the art SSMs organized according to the various paradigms proposed for their definition.

As a running example to illustrate the study, Fig. 1 presents a snapshot of the SNOMED-CT clinical healthcare terminology [18], in which biomedical concepts are organized by taxonomic relationships. The topology of SNOMED-CT defines a partial order $\preceq$ between concepts, e.g. '*Heparin*' $\preceq$ '*Protein*' means that the concept '*Heparin*' is subsumed by the concept '*Protein*', that is, the heparin is a specific class of protein.

2.1.1. Edge-based approaches

Edge-based measures estimate the similarity of two concepts according to the strength of their interlinking in the ontology. The most usual approach considers the similarity as a function of the distance which separates the two concepts in the ontology. For instance, Rada et al. estimate the distance of two concepts u, v as the shortest-path linking them ($sp(u, v)$) [15].

$$Dist_{Rada}(u, v) = sp(u, v) \tag{1}$$

In Fig. 1, the shortest path between the concepts c_5 and c_3 is $c_5 \rightarrow c_4 \rightarrow c_3$. Leacock and Chodorow proposed a non-linear adaptation of Rada's distance to define the similarity measure Sim_{LC} [19]:

$$Sim_{LC}(u, v) = -\log\left(\frac{sp(u, v)}{2 \cdot Max_depth}\right) \tag{2}$$

Rada's distance is here normalized by the maximal depth of the ontology, *Max_depth*, i.e. the longest of the shortest paths linking a concept to the concept which subsumes all the others (the root of the ontology, c_0 in Fig. 1).

More refined approaches propose to consider variations of the strength of the links between concepts; the deeper two linked concepts are, the stronger their semantic relationship will be

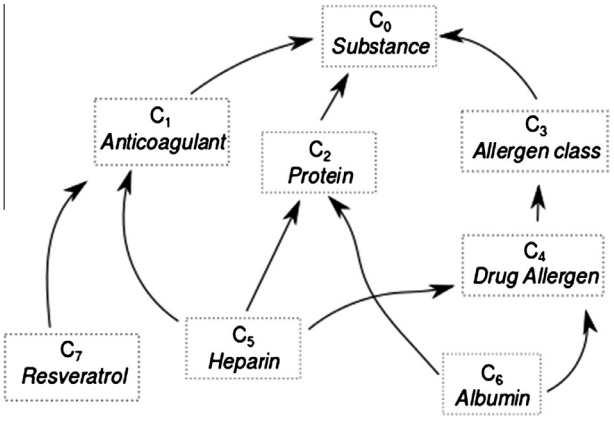


Fig. 1. Snapshot of the taxonomy of concepts defined in the SNOMED-CT.

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

considered. In most cases, the semantic similarity of two concepts is estimated as a function of the depth of the Least Common Ancestor (LCA), also named Least Common Subsumer (LCS), i.e. their common subsume, which has the maximum depth. In Fig. 1, the LCA of concepts c_5 and c_3 is c_0. The deepest the LCA, the more specific it is considered and, thus, the more similar the compared concepts are assumed. Based on this strategy, Wu and Palmer proposed Sim_{WP} [13] and Pekar and Staab proposed Sim_{PK} [20]:

$$Sim_{WP}(u,v) = \frac{2 \cdot depth(LCA_{u,v})}{depth(u) + depth(v)} \quad (3)$$

$$Sim_{PK}(u,v) = \frac{sp(LCA_{u,v}, root)}{sp(u, LCA_{u,v}) + sp(v, LCA_{u,v}) - sp(LCA_{u,v}, root)} \quad (4)$$

Most of these measures fulfil the *Identity Of the Indiscernibles* property (IOI), i.e. the similarity (resp. distance) of a concept to itself is maximal (minimal), e.g. $Dist_{Rada}(u,u) = 0$.

Table A1 in appendix presents some of the most usually referenced SSMs based on edge counting, some of their properties are also given.

2.1.2. Node-based approaches

Node-based approaches focus on the evaluation of concepts defined in the ontology. Two specific strategies can be distinguished: the feature-based and the one based on Information Theory.

2.1.2.1. Feature-based strategies. Feature-based strategies evaluate a concept as a set of features. These strategies are rooted into the *feature-model* proposed by Tversky [21]. For ontology-based measures, the features of a concept are usually considered as the set of concepts subsuming it, i.e. its ancestors $A(u) = \{v | u \preceq v\}$. In other words, a concept is characterized by the semantics that are inherited from its ancestors. In Fig. 1, the concept c_6 will therefore be represented by the set of features $A(c_6) = \{c_0, c_2, c_3, c_4, c_6\}$.

Feature-based strategies root semantic similarity in the context of classical binary or distance measures (e.g. set-based measures, vector-based measures). For example, Batet et al. assess taxonomic distance as the ratio between distinct and shared features [22]:

$$Dist_{Batet}(u,v) = \log_2\left(1 + \frac{|A(u) \setminus A(v)| + |A(v) \setminus A(u)|}{|A(u) \setminus A(v)| + |A(v) \setminus A(u)| + |A(u) \cap A(v)|}\right) \quad (5)$$

Another example of feature-based measure is given by Rodríguez and Egenhofer [23]:

$$Sim_{RE}(u,v) = \frac{|A(u) \cap A(v)|}{\gamma \cdot |A(u) \setminus A(v)| + (1-\gamma) \cdot |A(v) \setminus A(u)| + |A(u) \cap A(v)|} \quad (6)$$

with $\gamma \in [0,1]$, a parameter that enables to tune measure symmetry. Sim_{CMatch}, the Concept-Match similarity measure [24], is another example of feature-based similarity measure:

$$Sim_{CMatch}(u,v) = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|} \quad (7)$$

Some proposals adopting the feature-based strategy are presented in appendix Table A2.

2.1.2.2. Strategies based on Information Theory. Approaches based on Information Theory assess the similarity of concepts according to the amount of information they provide, i.e. their Information Content (IC). To this end, Resnik computes the IC of a concept according to Shannon's Information Theory as a function of its usage in a corpus [14]:

$$IC(u) = -log(p(u)) \quad (8)$$

with $p(u)$ the probability the concept u occurs in a document of the corpus. The more a concept occurs, the less informative it will be considered, assuming that a concept also occurs when the concepts it subsumes occur. By definition, the IC monotonically decreases from the terminal concept (leaves) to the root of the ontology. This ensures the taxonomic coherency of the results, i.e. $u \preceq v \Rightarrow IC(u) > IC(v)$. In other words, the IC of a concept will always be greater than the IC of one of its subsumers.

The similarity of two concepts is usually estimated according to the IC of their Most Informative Common Ancestor (MICA), i.e. the concept which subsumes the two concepts and has the maximum IC. In Fig. 1, concept c_1 is the MICA of the pair of concepts (c_5, c_7). As an example, Resnik proposes to estimate semantic similarity as follows [14]:

$$Sim_{Resnik}(u,v) = IC(MICA_{u,v}) \quad (9)$$

Resnik's measure does not explicitly capture the specificity, that is, the IC of evaluated concepts. Indeed, pairs of concepts with the same MICA will have identical similarity, even if, considering the taxonomical structure, their divergence towards their MICA is different. This is the case of the pairs (c_2, c_3) and (c_2, c_4), both having c_0 as MICA in Fig. 1. Therefore, Lin [25], Jiang and Conrath (JC) [12], Pirró and Euzenat (Sim_{Faith}) [26] and Mazandu and Mulder (Sim_{DIC}) [27], among others, e.g. [28], refined Resnik's measure to incorporate the IC of the compared concepts.

$$Sim_{Lin}(u,v) = \frac{2 \cdot IC(MICA_{u,v})}{IC(u) + IC(v)} \quad (10)$$

$$Dist_{JC}(u,v) = IC(u) + IC(v) - 2 \cdot IC(MICA_{u,v}) \quad (11)$$

$$Sim_{Faith}(u,v) = \frac{IC(MICA_{u,v})}{IC(u) + IC(v) - IC(MICA_{u,v})} \quad (12)$$

$$Sim_{DIC}(u,v) = \frac{2\sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c) + \sum_{c \in A(v)} IC(c)} \quad (13)$$

Note that the above measures only consider the MICA to estimate the information shared by two concepts. Indeed, in most cases, the MICA summarizes the information contained in the set of shared ancestors, as it is subsumed by this whole set. However, in some cases, due to multiple inheritances, the notion of MICA only captures the information shared by two concepts partially. For instance, in Fig. 1, considering that $IC(c_4) > IC(c_2)$, the MICA of concepts c_5 and c_6 is c_4. However, the amount of information shared by c_5 and c_6 is composed of the amount of information carried by $\{c_4, c_2\}$, their Set of Least Common Ancestors (SLCAs). In this case, MICA-based measures will only consider the most informative concept shared by two concepts. To better estimate the information shared by two concepts, Couto et al. proposed GraSM and DiShIn strategies in which the IC of the SLCAs of two concepts are aggregated [29,30].

In appendix, Table A3 presents some well-known SSMs based on Information Theory.

The cornerstone of the above measures is the accurate estimation of the IC of concepts. In order to avoid depending on annotated corpora, whose creation is time consuming, and which are sometimes difficult to obtain (due to data sensibility, e.g. patient record) [31], various *intrinsic* IC calculus models have been proposed. They estimate the IC of concepts by only considering structural information extracted from the ontology. Intrinsic IC calculus can be based on multiple topological characteristics such as the number of descendants, ancestors and depth [31–33]. Seco et al. [32] propose computing the IC of a concept as a function of its number of descendants:

$$IC_{Seco} = 1 - \frac{\log(D(u))}{\log(|C|)} \quad (14)$$

with $D(u) = \{v | v \preceq u\}$ and C the set of concepts defined in the ontology.

In another approach, Sánchez et al. [31] estimate the IC of a concept according to the ratio between the number of terminal concepts (leaves) it subsumes and the amount of ancestors it has:

$$IC_{Sanchez}(u) = -\log\left(\frac{\frac{|leaves(u)|}{|A(u)|} + 1}{Max_Leaves + 1}\right) \quad (15)$$

with $leaves(u)$ as the number of leaves subsumed by the concept u (e.g. $leaves(c_2) = \{c_5, c_6\}$ in Fig. 1) and Max_Leaves as the number of terminal concepts of the ontology.

2.1.3. Hybrid approaches

Hybrid approaches combine notions from edge-based and node-based approaches [34–37]. They are usually defined as ad hoc and weighted aggregations of ancestors, node degrees and concept specificities (e.g. IC) [35]. In appendix, Table A4 presents some proposals based on this principle.

2.2. Related work on unifying semantic measures

Tversky was the first to formulate a framework of semantic similarity, the *feature model*, from which a family of semantic measures can be derived [21]. The feature model proposes to characterize semantic objects in a broad sense and was not originally defined for ontological concepts. This model requires the semantic objects to be represented as sets of features. Their similarity is therefore intuitively defined as a function of their common and distinctive features, an approach commonly used to compare sets (e.g. Jaccard index). Tversky defined two parameterized SSMs: the *contrast model* (Sim_{CM}) and the *ratio model* (Sim_{RM}). They can be used to compare two semantic objects (u, v) through their respective sets of features U and V:

$$Sim_{CM}(u, v) = \gamma f(U \cap V) - \alpha f(U \setminus V) - \beta f(V \setminus U) \quad (16)$$

$$Sim_{RM}(u, v) = \frac{f(U \cap V)}{\alpha f(U \setminus V) + \beta f(V \setminus U) + f(U \cap V)} \quad (17)$$

with α, β and $\gamma \geqslant 0$.

Note that, considering f as the cardinality of the set, setting Sim_{RM} with $\alpha = \beta = 1$ leads to the Jaccard index, and setting $\alpha = \beta = 0.5$ leads to the Dice coefficient. In other words, set-based measures can be used to easily express abstract formulations of similarity measures.

The framework proposed by Tversky "just" defines that a semantic object can be represented as a set of features and that commonalities and differences must be evaluated to assess the similarity. By definition, Sim_{CM} and Sim_{RM} are therefore constrained in the set-based frame, i.e. they require compared objects to be represented as sets. They are, however, considered abstract measures as they rely on an undefined function f, specifying how to capture the sets of features of compared objects.

Among the large diversity of proposals, most set-based measures can be split into two groups: those based on the Caillez and Kuntz σ_α formulation, and those based on Gower and Legendre σ_β formulation [17]. Since set-based measures can be used to design semantic measures, σ_α and σ_β can be expressed in a straightforward manner according to the Tversky feature approach.

$$\sigma_\alpha(u, v) = \frac{f(U \cap V)}{\left(\frac{f(U)^\alpha + f(V)^\alpha}{2}\right)^{1/2}} \quad (18)$$

$$\sigma_\beta(u, v) = \frac{\beta f(U \cap V)}{f(U) + f(V) + (\beta - 2) f(U \cap V)} \quad (19)$$

The abstract formulation σ_α can be used to express, among others, Simpson ($\alpha = -\infty$) and Ochaiai ($\alpha = 0$) coefficients [38]. The σ_β reformulation enables the definition of other numerous measures, e.g. Sokal and Sneath ($\beta = 0.5$), and Jaccard ($\beta = 1$) and Dice ($\beta = 2$) coefficients [17,38].

Blanchard et al. were the first to take advantage, in an explicit manner, of abstract SSMs to compare pairs of concepts defined in an ontology [17]. They focus on an information theoretic semantic similarity to underline relationships between several existing measures. For example, based on the intuitive notion of commonality and differences, they underlined that Wu & Palmer and Lin similarity measures, Eqs. (3) and (10), can be derived from the Dice index. They were also the first to stress the suitability of an abstract framework to define new measures and to study properties of groups of measures [17].

Other authors also demonstrated the relationships between, a priori, different similarity measures and took further advantage of frameworks to design new measures [8,26,39-41]. These contributions mainly focused on establishing local relationships between set-based measures and measures based on Information Theory. As an example, Sánchez and Batet [8] proposed a framework, grounded in Information Theory, which allows several measures (i.e., edge-counting and set-based coefficients) to be uniformly redefined according to the notion of IC. Cross et al. also proposed a similar contribution in which feature-based approaches and measures based on Information Theory are expressed through the frame of the fuzzy-sets theory [40–42].

Despite the suitability of existing frameworks for studying some of the SSM properties, only a few works rely on them to express measures [8,40]. Moreover, current frameworks only focus on a specific paradigm (e.g. feature-based strategy), which is used to express SSMs. In fact, existing frameworks only encompass a limited number of measures and were not defined in the purpose of unifying SSMs expressed using the variety of paradigms reviewed in Section 2. The following section is dedicated to the definition of such a unifying theoretical framework.

3. A unifying framework for ontology-based semantic similarity measures

The analysis of the state-of-the-art allowed us to distinguish a few core elements underlying most SSMs. Their notation and meaning are given in this section. The abstract measures, which can be defined as a function of those core elements, are then introduced and discussed. Finally, we illustrate the suitability of the proposed framework to express a selection of well-known SSMs available in the literature.

3.1. Notation

In order to ease the readability of this section, the various notations which will be used to present the framework are listed below; provided examples are presented in association with Fig. 2:

- G: the semantic graph (ontology) in which concepts are defined.
- C: the set of concepts in, e.g., $C = \{c_0, c_1, \ldots, c_9\}$.
- $A(u)$: the set of concepts including u and its ancestors, defined by the partial order $\preceq$ of G, i.e. $A(u) = \{v \in C | u \preceq v\}$. For example, $A(c_8) = \{c_8, c_1, c_2, c_0\}$.
- $D(u)$: the set of descendants of c, i.e. $D(u) = \{v \in C | v \preceq u\}$. For example, $D(c_8) = \{c_8, c_9\}$.

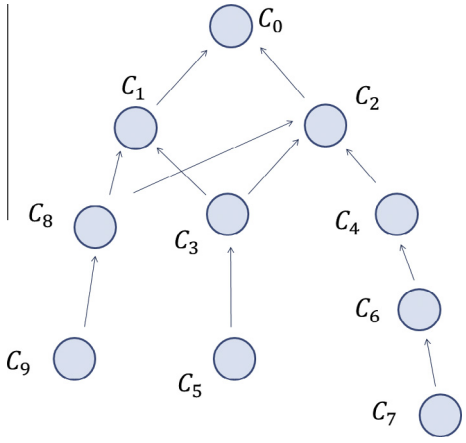


Fig. 2. Taxonomical semantic graph defining a partial ordering between a set of concepts.

- *leaves*(C): the concepts respecting $D(c) = \{c\}$, *e.g.* *leaves*(C) = $\{c_9, c_5, c_7\}$.
- $\{u \rightsquigarrow v\}$: the set of taxonomic paths leading from concept u to concept v, e.g., $\{c_8 \rightsquigarrow c_0\} = \{c_8 \rightarrow c_1 \rightarrow c_0, c_8 \rightarrow c_1 \rightarrow c_3 \rightarrow c_2 \rightarrow c_0, \ldots\}$.
- $\Omega_{u,v} \subseteq A(u) \cap A(v)$: the Set of Least Common Ancestors (SLCAs) of concepts u and v, i.e., the *minimal* set of shared ancestors of u and v which are subsumed by the maximal number of common ancestors of the two concepts. The idea is to distinguish the set of concepts which contains all the meaning carried by the concepts subsuming u and v. $\Omega_{u,v}$ therefore corresponds to the set of ancestors of both u and v which are the more specific (e.g., deeper in the ontology) and non-comparable with regard to the partial order defined in G, i.e., $x \not\leqslant y$ and $y \not\leqslant x$. Otherwise stated, $\Omega_{u,v}$ corresponds to the leaves of the graph induced by the concepts found in $A(u) \cap A(v)$. For example, in Fig. 2, we obtain $\Omega_{c_9,c_5} = \{c_1, c_2\}$ and $\Omega_{c_7,c_5} = \{c_2\}$. More formally, the SLCAs of the concepts u and v can be defined as the minimal set of concepts respecting $\bigcup_{c \in \Omega_{u,v}} A(c) = A(u) \cap A(v)$. The notion of SLCAs have also been introduced through the term Disjoint Common Ancestors (DCAs) in the literature [29]. We abbreviate $\Omega_{u,v}$ by Ω when there is no ambiguity.
- $G_c \subseteq G$: the graph induced by a concept c, considering both its ancestors and descendants.
- G_c^+ (resp.G_c^-): a graph induced by a concept C, only considering its ancestors (resp. descendants). For instance, $G_{C_3}^+$ is the sub-graph of G, only considering concepts $A(c_3) = \{c_0, c_1, c_2, c_0\}$ and associated relationships.
- $\mathbb{K}$: a domain containing any subset or subgraph of G: $\{u \rightsquigarrow v\}, A(u), D(u), G_u, G_u^+$.

3.2. Core elements of semantic similarity measures

In this section, we first distinguish the core elements of SSMs; secondly, we will further detail each of them through concrete examples.

As stated in Section 2.1, similarity measures are designed according to specific paradigms (e.g., edge-counting or node-based strategies). Therefore, measure designers first adopt a specific paradigm from which estimators of commonalities and differences will be defined. They next adopt a strategy by which those estimators will be aggregated to express a similarity measure or a taxonomical distance. Indeed, in a broad sense, when comparing two *things*, their commonalities and differences are the only evidences from which similarity (or dissimilarity) can be evaluated. In the aim of distinguishing the core elements of SSMs, estimators of commonalities and differences intuitively appear as critical elements of semantic measures. In fact, they are the roots of *all* existing similarity measures.

The definition of the estimators of commonalities and differences depends on the paradigm chosen to formulate SSMs. For instance, for edge-counting approaches, the difference of two concepts is assessed as a function of the length of the shortest path linking them, while for feature-based approaches, concept differences are computed as a function of the features characterizing a concept, which are not shared with the other.

The main differences between existing paradigms depend on the strategy adopted to represent a concept. Such representation will determine the expressions of the estimators of commonalities and differences. Therefore, we formally introduce a function aiming at representing a concept.

Definition 1. The mapping of a set of concept $C' \subseteq C$ to its *semantic representation*, denoted $\widetilde{C'}$, which encompasses its semantic features, is defined by the function $\rho(C')$:

$$\rho : P(C) \rightarrow \mathbb{K}$$

For convenience, we note $\rho(u)$ and $\tilde{u}$, the representation of a single concept u, i.e.$\{u\}$. Concrete examples of the core elements will be discussed later in this section.

We also formally define the function aiming to estimate the commonalities and differences of two concepts (u, v), according to their semantic representations $(\tilde{u}, \tilde{u})$:

Definition 2. The commonality of two concept representations $(\tilde{u}, \tilde{u})$ is estimated using a function $\Psi(\tilde{u}, \tilde{u})$:

$$\Psi : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$$

Definition 3. The amount of knowledge represented in $\tilde{u}$ not found in $\tilde{v}$ is estimated using a function $\Phi(\tilde{u}, \tilde{u})$:

$$\Phi : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$$

Those three abstract functions ρ, Ψ, Φ are the core elements of most similarity measures. In the context of semantic similarity estimation, they can be used to reformulate, in an abstract manner, all SSMs based on commonalities and differences of compared concepts. As an example, an estimation of the shortest-path linking two concepts u, v [15] can be abstracted to the sum of their differences, being estimated regarding their LCA:

$$sp(u, v) \approx \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u})$$

where $\Phi(\tilde{u}, \tilde{v}) = sp(u, LCA_{u,v})$ and $\Phi(\tilde{v}, \tilde{u}) = sp(v, LCA_{u,v})$,with $LCA_{u,v} \in sp(u, v)$

Designers of similarity measures sometimes consider the whole semantic space in which compared elements are modelled [38]. We therefore define a function enabling to capture this information.

Definition 4. The amount of knowledge defined in G (i.e. modelled in the ontology), which is neither found in $\tilde{u}$ nor in $\tilde{v}$, can be estimated by a function $\zeta(\tilde{u}, \tilde{v})$:

$$\zeta : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$$

Fig. 3 presents an intuitive feature-based representation of the functions introduced by the framework. The representation of the concept u, i.e. $\rho(u)$, is here defined as $A(u)$, i.e. the set of subsumers of u. The commonalities and differences (Ψ, Φ) of two concept representations are intuitively defined by the set operators ($\cap$ and

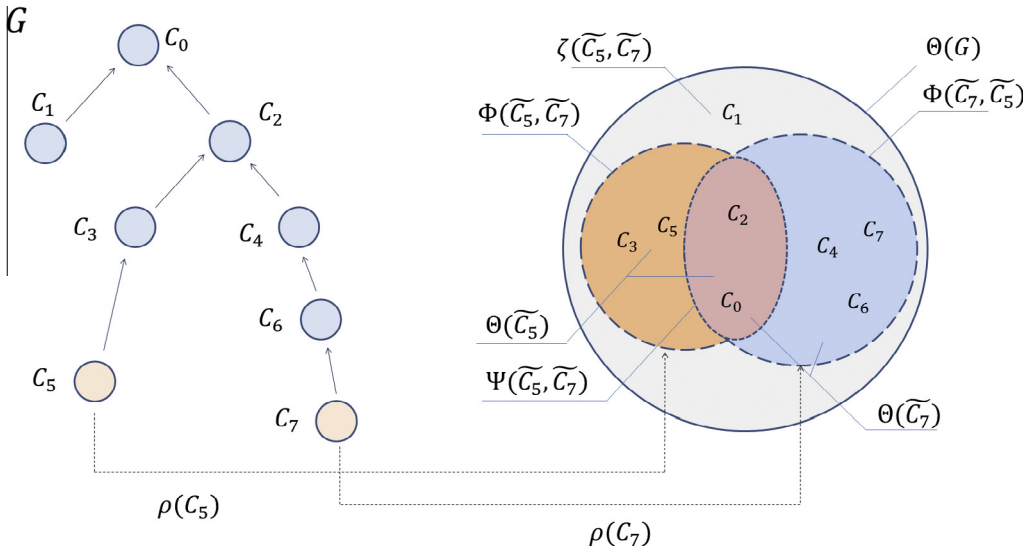


Fig. 3. Example of expression of the framework's core elements according to the feature-based approach.

\ respectively). The part of the universe which is not contained in compared concepts is denoted by ζ.

Most measures can be expressed in an abstract manner using the functions ρ, Ψ, Φ and, in some particular cases, ζ. However, there are situations in which functions Ψ and Φ may also be expressed according to the specificity of a concept or, more generally, according to the amount of information carried by a concept representation (e.g. Lin and Resnik for information theoretic measures). Thus, we further define two functions capturing these notions.

Definition 5. The specificity of a concept u is estimated by a function $\theta(u)$.

$$\theta : C \rightarrow \mathbb{R}^+$$

The expressions used to compute the IC of a concept, presented in Section 2.1.2.2, are particular expressions of function θ. Other estimators, which are not framed in Information Theory (e.g. the depth of a concept) can be used to estimate concepts' specificity.

Finally, we also generalize the notion of specificity of a concept to a representation of a concept:

Definition 6. The degree of specificity of a concept representation $\tilde{u}$ can be estimated by a function $\Theta(\tilde{u})$:

$$\Theta : \mathbb{K} \rightarrow \mathbb{R}^+$$

The Θ function generalizes function θ defined to estimate concept specificity. This is required to express state-of-the-art semantic measures based on an aggregation of θ [27,43]. As an example, considering the representation of concept u by $\tilde{u} = A(u)$, in Eq. (13), Mazandu et al. defined $\Theta(\tilde{u})$ as:

$$\Theta(A(u)) = \sum_{c \in A(u)} \theta(u)$$

We further detail how the various core elements distinguished by the framework can be associated to ontological knowledge.

3.2.1. Mapping a concept to its semantic representation in the ontology (ρ)

The semantic representation of a set of concepts contained in an ontology can be viewed as a subset of the knowledge that the ontology models. Thus, the function ρ defines the mapping between a set of concepts and its semantic representation in the ontology. We first consider the case in which the set of concepts only contains a single concept. Fig. 4 shows some semantic representations of a concept that are commonly used to design semantic measures.

One of the most general semantic representations of a concept u is G_u, i.e. the graph induced by the ancestors ($A(u)$) and the descendants ($D(u)$) of u. However, in most cases, SSMs are based on G_u^+, the graph induced by u, only considering its ancestors Indeed, as stressed in Fig. 4, from G_u^+, multiple concept representations can be derived, such as the set of ancestors $A(u)$ or the set of paths linking the concept to the root $\{u \rightsquigarrow root\}$ to cite a few. As we have seen in Section 2.1.2.1, representing a concept by $A(u)$ is extensively used to express measures based on the feature approaches [22,23], or based on Information Theory [12,25,44]. Moreover, the representation of a concept through $\{u \rightsquigarrow root\}$ is commonly adopted in defining measures based on the edge-counting approach [13,15,20].

In order that the function ρ is defined for a set of concepts, we consider that union operators are defined for the proposed concept representations. This is indeed the case of all representations based on sets and of those corresponding to graphs. Formally, the representation of a set of concepts $C' \subseteq C$ can be derived from the representation of a single concept, i.e. $\rho(C') = \bigcup_{u \in C'} \rho(u)$, e.g. defining $\rho(C') = \bigcup_{u \in C'} A(u)$.

3.2.2. Estimating concept specificity (θ and Θ)

Numerous measures rely on the notion of the amount of information captured by a concept. The notion of Information Content (IC) exploited by information theoretic measures was defined for this purpose. Other strategies, which are not grounded in Information Theory, have also proposed to evaluate the specificity of a concept according to, for instance, its depth in the ontology [13]. We therefore generalized the notion of IC by introducing a function θ which estimates the specificity of a concept. Since the central element of the framework is the semantic representation of a concept (ρ), we also introduced a function Θ, which assesses the specificity of that semantic representation; in other words, this function generalizes θ. In the same manner as θ, and in coherency with the taxonomical structure, Θ also decreases monotonically from the leaves to the root of the ontology, i.e. $u \preceq v \Rightarrow \Theta(\tilde{u}) > \Theta(\tilde{v})$.

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

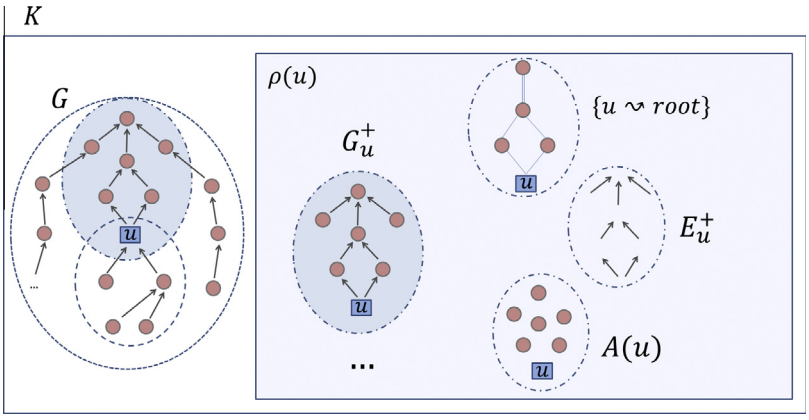


Fig. 4. Semantic representations of a concept commonly used to design semantic similarity measures.

Depending on the representation adopted for ρ, various strategies can be defined to evaluate $\Theta(\tilde{u})$. Without loss of generality, we focus here on the case where $\Theta(\tilde{u})$ is assessed for $\tilde{u} \subseteq G_u^+$, e.g. $\tilde{u} = A(u)$. Two commonly used strategies are briefly discussed:

- Evaluating the cardinality of $\tilde{u}$. We obtain $\Theta(\tilde{u}) = |A(u)|$, which can be substituted by $\theta(u)$ so that $\theta(u) = |A(u)|$. In this case, a commonly used strategy is to define $\Theta(\tilde{u}) = max_{c \in A(u)}\theta(\tilde{c}) = \theta(u)$. This strategy was adopted by Lin, Resnik, Wu, Palmer and numerous SSMs.
- Aggregating the specificity of the concept contained in $A(u)$ by considering a particular θ function. This leads to $\Theta(\tilde{u}) = \sum_{c \in A(u)}\theta(c)$, which can be extended to $\Theta(\tilde{u}) = \sum_{c \in A(u)}\Theta(\tilde{c})$. Mazandu and Mulder recently proposed an information content using such strategy to evaluate the specificity of a concept [27] (see Eq. (13)).

3.2.3. Estimating the commonalities of two concepts (Ψ)

The commonality between concept pairs, evaluated by the function Ψ, can be associated with the amount of information captured by features *shared* among the semantic representations of these concepts, i.e. intuitively $\Theta(\tilde{u} \cap \tilde{v})$. For example, when $\tilde{u}$ is associated with sets, e.g. $\tilde{u} = A(u)$, a commonly used strategy is to define commonalities by $|A(u) \cap A(v)|$ (e.g. Eq. (6), Sim_{RE}). In other words, the function Ψ assesses the specificity of the features shared between the semantic representations of the compared concepts.

Numerous similarity measures consider taxonomies as tree structures (e.g. [13,20]). In a tree, there is just a single concept ω that subsumes two other concepts u, v such that $A(\omega) = A(u) \cap A(v)$. The notions of LCA/LCS and MICA correspond to this concept ω. Moreover, in a tree, the depth of a concept is particularly suited to evaluate its specificity as $A(u) = 1 + depth(u)$. Thus, in trees, the function Ψ can assess the commonalities of two concepts by just considering $\Theta(\tilde{\omega})$, defining, for example, $\Theta(\widetilde{LCA}_{u,v})$. However, because of the presence of multiple taxonomic inheritances in most widely used ontologies (e.g., SNOMED-CT, MeSH), the notion of a single subsuming concept ω characterizing the whole commonality of two concepts is not usually fulfilled. Therefore, in order to capture commonalities, $\Psi(\tilde{u}, \tilde{v})$ must define an aggregation strategy while taking into account the specificity of all concepts composing $\Omega_{u,v}$, that is, the set of non-comparable common ancestors of concepts u and v. In other words, for most ontologies, ω must be generalized to $\Omega_{u,v}$.

Each concept in $\Omega_{u,v}$ represents a particular semantic facet of the commonality between the concepts u and v. Some approaches which evaluate these commonalities explicitly aggregate the amount of information of the semantic facets in Ω [14,25]. However, most measures adopt the maximal strategy as they only exploit ω^*, that is, the concept of Ω which maximizes a selected θ function. Measures relying on the MICA (e.g. Lin [25], Resnik [14]) or on the LCA (e.g. Wu and Palmer [13]) are examples of this strategy. Nevertheless, other aggregations have been proposed [29,30]; for example, GraSM propose to average the specificities of concepts in Ω:

$$\Psi_{\text{GraSM}}(\tilde{u}, \tilde{v}) = \frac{\sum_{c \in G_\Omega^+}\theta(c)}{|\Omega|}$$

Note that for ontologies incorporating multiple inheritances, the commonality of a pair of concepts can be also estimated by taking into account their common descendants, which can be seen as their shared potential extensions. The problem is symmetrical to the estimation of the commonality based on the shared ancestors Ω (which could be renamed Ω^+). Likewise, a set Ω^- representing the non-comparable common descendants of two concepts can also be expressed. Estimation of concepts' commonality based on the study of their descendants has been recently introduced in [45].

As we have seen, evaluating the commonalities of two terms is equivalent to evaluating the specificity of the semantic representation built from the group of concepts Ω. Existing approaches (LCA/MICA [13,14], GraSM and DiShIn [29,30]) only aggregate the specificity of the semantic facets represented by Ω.

3.2.4. Estimating the differences of two concepts (Φ)

Some measures also rely on the differences between the semantic representations of the compared concepts, to which we refer as function Φ. Considering two concepts u,v, the information contained in u that is not in v is intuitively expressed by:

$$\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v}) \tag{20}$$

In practice, Φ is usually computed as $\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Theta(\widetilde{\Omega})$, with $\Theta(\widetilde{\Omega})$ as the amount of information carried by the set of concepts in Ω. Moreover, similarly to Ψ, numerous Φ approaches only consider ω^*, i.e. the concept from Ω, maximizing an expression of the function θ. This results in $\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Theta(\tilde{\omega}^*)$, which is usually expressed by $\Phi(\tilde{u}, \tilde{v}) = \theta(u) - \theta(\omega^*)$ (e.g. [13,14,46]). We present an example of such a formulation used in the well-known Jiang and Conrath measure [12]:

$$\begin{aligned} Dist_{JC}(u, v) &= IC(u) + IC(v) - 2 \cdot IC(MICA_{u,v}) \\ &\approx \Theta(\tilde{u}) + \Theta(\tilde{v}) - 2 \cdot \Psi(\tilde{u}, \tilde{v}) \\ &\approx \Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v}) + \Theta(\tilde{v}) - \Psi(\tilde{u}, \tilde{v}) \\ &\approx \theta(u) - \theta(\omega^*) + \theta(v) - \theta(\omega^*) \approx \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u}) \end{aligned}$$

Thus, by defining $\Phi(\tilde{u}, \tilde{v}) = IC(u) - IC(MICA_{u,v})$, we obtain:

$$Dist_{JC}(u, v) = IC(u) + IC(v) - 2 \cdot IC(MICA_{u,v})$$

For edge-counting approaches, as introduced in Section 3.2, the differences of a concept u with respect to v are usually assessed from the length of the shortest path between the concept and their LCA. Thus:

$$\Phi(\tilde{u}, \tilde{v}) = sp(u, LCA_{u,v})$$

Other strategies can be defined to aggregate the differences between the concept and those contained in Ω. As an example, some node-based measures (e.g. Sim_{DIC} Eq. (13)) take into account all the information related to Ω, as follows:

$$\Phi(\tilde{u}, \tilde{v}) = \Theta\left(A(u) \setminus \bigcup_{c \in \Omega} A(c)\right) = \Theta(A(u) \setminus (A(u) \cap A(v)))$$

Despite particular instantiations of Φ, the vast majority of measures exploiting semantic differences (e.g. Jiang and Conrath, Lin, Resnik, Rada, Wu and Palmer, see equations Section 2) estimate the dissimilarity between two concepts as $\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$.

3.3. Expressing semantic measures

Once we have distinguished the core elements of SSMs, they can be used to express a large diversity of measures by instantiating abstract expressions such as set-based coefficients. To illustrate the generality and possibilities of the proposed framework, we present some instantiations corresponding to existing measures that can be obtained from the Jaccard index. The Jaccard index can be intuitively generalized by considering an abstraction of the set-based operators, i.e., with U, V two sets, $\Psi(\tilde{u}, \tilde{v}) = U \cap V$ and $\Phi(\tilde{u}, \tilde{v}) = U \setminus V$, as follows:

$$Sim_{Jaccard}(u, v) = \frac{\Psi(\tilde{u}, \tilde{v})}{\Psi(\tilde{u}, \tilde{v}) + \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u})}$$

considering $\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$.

$$Sim_{Jaccard}(u, v) = \frac{\Psi(\tilde{u}, \tilde{v})}{\Theta(\tilde{u}) + \Theta(\tilde{v}) - \Psi(\tilde{u}, \tilde{v})}$$

Based on specific expressions of the functions Ψ and Φ (see Table 1), the Jaccard index can be used to express Pekar and Staab (Sim_{PK}) [20], Sim_{Faith} [26] or Sim_{CMatch} [24] (see Eqs. (4), (12), (7), Section 2.1). It can also be used to express Sim_{cGIC}, an unpublished pairwise measure based on Sim_{GIC} [43], which was initially designed to compare groups of concepts:

$$Sim_{cGIC}(u, v) = \frac{\sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c) + \sum_{c \in A(v)} IC(c) - \sum_{c \in A(u) \cap A(v)} IC(c)}$$

The proposed framework also enables taking advantage of studies made for other binary measures (e.g. measures used to compare vectors or sets). Indeed, the proposed core elements can be mapped to existing theoretical tools used by other communities for studying binary measures. Table 2 shows abstract expressions of the Operational Taxonomic Units (OTUs), classically used to represent binary measures [38]. Thus, such mapping can be used to easily express SSMs based on binary measure expressions relying on OTUs [38].

Table 2
Links between Operation Taxonomic Units (OTUs) commonly used for the definition of binary measures and the theoretical framework core elements (see Choi et al. [38] for numerous expressions of binary measures using OTUs).

$u \setminus v$	$\Theta(\tilde{v})$	$\overline{\Theta(\tilde{v})}$
$\Theta(\tilde{u})$	$\Psi(\tilde{u}, \tilde{v})$	$\phi(\tilde{u}, \tilde{v})$
$\overline{\Theta(\tilde{u})}$	$\phi(\tilde{v}, \tilde{u})$	$\zeta(\tilde{u}, \tilde{v})$

In a similar manner to the approach relying on information theory, the amount of information expressed in an ontology G can be viewed as $\Theta(G)$. The amount of information encompassed in the semantic representation of a concept is expressed by $\Theta(\tilde{c})$, and the amount of information expressed in G, which is not found in c can be defined by $\overline{\Theta(\tilde{c})}$.

Therefore, based on the seventy expressions of binary measures distinguished by Choi et al. [38] and the correspondences proposed in Table 2, numerous new measures can easily be expressed. The main idea is to generalize existing binary measures using the proposed core elements of the framework to derive numerous SSMs, as performed above using an abstract formulation of the Jaccard index. Three examples of measures expressions are presented:

$$Sim_{Dice}(u, v) = \frac{2\Psi(\tilde{u}, \tilde{v})}{2\Psi(\tilde{u}, \tilde{v}) + \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u})}$$

$$Dist_{Hamming}(u, v) = \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u})$$

$$Sim_{Braun-Blanquet}(u, v) = \frac{\Psi(\tilde{u}, \tilde{v})}{\max(\Psi(\tilde{u}, \tilde{v}) + \Phi(\tilde{u}, \tilde{v}), \Psi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u}))}$$

This section illustrates that distinguishing the core elements of SSMs and defining them at an abstract level is important to highlight relationships between them, and to better understand and capture the semantics associated to a measure. Moreover, based on contributions made for other types of similarity measures, we explicitly established a relationship between existing theoretical tools and show how a large diversity of SSMs can be easily generated.

3.4. Unification of abstract measures

In this sub-section, we demonstrate the relationships between known abstract expressions of measures through the definition of a new parameterized SSM.

In previous sections we have identified the core elements of SSMs and we have shown how they can be used to instantiate and design specific measures. Moreover, we have underlined that set-based measures can be used to express abstract measures. By extension, Caillez and Kuntz σ_α and Gower and Legendre σ_β formulas (presented in Section 2.2) may be considered abstract parameterized measures. Moreover, by focusing on the unification of measure expressions, we here demonstrate that under some conditions, σ_α and σ_β can be partially unified and extended through a common expression.

We first demonstrate that σ_α can be easily extended to the well-known generalized mean of order α (Result 1). In addition, we show that σ_β is a particular case of the *ratio model* proposed by Tversky (Result 2). Finally, based on Results 1 and 2, we demonstrate that a new abstract tunable measure can be used to express a large diversity of abstract measure (Result 3).

Table 1
Examples of particular expressions of core elements, from which similarity measures can be obtained as instantiations of the Jaccard index. These can also be used to obtain other measures using different set-based coefficients.

Core elements	Sim_{PK}	Sim_{Faith}	Sim_{CMatch}	Sim_{cGIC}
$\rho(u) = \tilde{u}$	G_u^+	$A(u)$	$A(u)$	$A(u)$
$\Theta(\tilde{u})$	$sp(u, root)$	$IC(u)$	$\lvert A(u)\rvert$	$\sum_{c \in A(u)} IC(c)$
$\Psi(\tilde{u}, \tilde{u})$	$sp(LCA_{u,v}, root)$	$IC(MICA_{u,v})$	$\lvert A(u) \cap A(v)\rvert$	$\sum_{c \in A(u) \cap A(v)} IC(c)$
$\Phi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

Result 1. *First, note that Cauchy's mean σ_α implies a symmetric contribution of $\Theta(\tilde{u})$ and $\Theta(\tilde{v})$. In a straightforward manner, we extend σ_α to the generalized mean of order α [47] by introducing two parameters x and y, enabling us to tune $\Theta(\tilde{u})$ and $\Theta(\tilde{v})$ contributions.*

$$\sigma_{\alpha,x,y}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{\left(x\cdot\Theta(\tilde{u})^\alpha + y\cdot\Theta(\tilde{v})^\alpha\right)^{1/\alpha}} \tag{21}$$

With $x+y=1$ and $x,y\geqslant 0$. σ_α is a special case of $\sigma_{a,x,y}$ when $x=y=\frac{1}{2}$.

Result 2. *We demonstrate the relationship between σ_β and the ratio model* (Eqs. (19) and (17) *respectively). Recall that $\Theta\tilde{u}$ (resp. $\Theta\tilde{v}$) represents the specificity of the representation of a concept u (resp. v). σ_β is defined by*:

$$\sigma_\beta(u,v) = \frac{\beta\Psi(\tilde{u},\tilde{v})}{\Theta(\tilde{u})+\Theta(\tilde{v})+(\beta-2)\Psi(\tilde{u},\tilde{v})}$$

Note that the function $\Theta_{\tilde{u}}$ is commonly considered as additive, i.e. $\Theta(\tilde{u}\cup\tilde{v}) = \Theta(\tilde{u})+\Theta(\tilde{v})$ for any pair of non-comparable semantic representations $(\tilde{u},\tilde{v})$. With this condition we can demonstrate the following lemma.

Lemma. *Considering $\Theta(\tilde{u}) = \Phi(\tilde{u},\tilde{v})+\Psi(\tilde{u},\tilde{v})$, σ_β (Gower and Legendre abstract formulation) is a particular case of the ratio model.*

Proof. Considering the inverse of both σ_β and the *ratio model* Sim_{RM}, we obtain: □

$$Sim_{RM}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{x\Phi(\tilde{u},\tilde{v})+y\Phi(\tilde{v},\tilde{u})+\Psi(\tilde{u},\tilde{v})}$$

Setting $x=y$, we have:

$$\frac{1}{Sim_{RM}(u,v)} = 1 + x\frac{\Phi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})} + x\frac{\Phi(\tilde{v},\tilde{u})}{\Psi(\tilde{u},\tilde{v})}$$

In addition,

$$\begin{aligned}\frac{1}{\sigma_\beta(u,v)} &= 1-\frac{2}{\beta}+\frac{1}{\beta}\frac{\Theta(\tilde{u})}{\Psi(\tilde{u},\tilde{v})}+\frac{1}{\beta}\frac{\Theta(\tilde{v})}{\Psi(\tilde{u},\tilde{v})}\\ &= 1-2x+x\frac{\Theta(\tilde{u})}{\Psi(\tilde{u},\tilde{v})}+x\frac{\Theta(\tilde{v})}{\Psi(\tilde{u},\tilde{v})}\end{aligned}$$

Considering $\Theta(\tilde{u}) = \Phi(\tilde{u},\tilde{v})+\Psi(\tilde{u},\tilde{v})$, we obtain:

$$\begin{aligned}\frac{1}{\sigma_\beta(u,v)} &= 1-2x+x\frac{\Phi(\tilde{u},\tilde{v})+\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})}+x\frac{\Phi(\tilde{v},\tilde{u})+\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})}\\ &= 1+x\frac{\Phi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})}+x\frac{\Phi(\tilde{v},\tilde{u})}{\Psi(\tilde{u},\tilde{v})} = \frac{1}{Sim_{RM}(u,v)}\end{aligned}$$

Thus, σ_β is a particular case of the *ratio model* proposed by Tversky, considering an equal contribution of $\Theta(\tilde{u},\tilde{v})$ and $\Phi(\tilde{u},\tilde{v})$ (i.e. $x=y$).

Result 3. *$\sigma_{\alpha,x,y}$ and the ratio model (which includes, see result 2) may be expressed by the general function $\Sigma_{\alpha,x,y,z}$ (shorten by Σ).*

$$\Sigma(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{\left(x\cdot\Theta(\tilde{u})^\alpha + y\cdot\Theta(\tilde{v})^\alpha + z\cdot\Psi(\tilde{u},\tilde{v})^\alpha\right)^{1/\alpha}} \tag{22}$$

with $x,y,z\geqslant 0$ and $x+y+z=1$. Note that by setting $\alpha=1$ and $\Theta(\tilde{u}) = \Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}))$ the abstract measure Σ can also be formulated as:

$$\Sigma(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{x\cdot\Phi(\tilde{u},\tilde{v})+y\cdot\Phi(\tilde{v},\tilde{u})+(x+y+z)\cdot\Psi(\tilde{u},\tilde{v})} \tag{23}$$

In this section we have demonstrated that existing abstract measures can be generalized to the Σ abstract measure, from which a large diversity of measures can be derived. Unifying abstract measures opens interesting perspectives for measure optimization. Indeed, expressing existing measures through a common parameterized formula enable better understanding the relationships between the various proposals. Moreover, a large variety of measures can easily be instantiated by tuning few parameters. Unification of measures is therefore a prerequisite in order to distinguish the parameters best impacting measure accuracy.

4. Practical application of the framework

This section is devoted to the application of the proposed framework in a practical setting. The main goal is to discuss how SSMs can be designed and studied by means of the core elements identified by the framework, from which the most commonly used are:

- The estimator of the specificity of a concept defined in an ontology (function θ).
- The representation of a concept or a set of concepts corresponding to a canonical form which can be processed in order to extract the semantics of the (set of) concept(s) (function ρ).
- The estimator of the specificity of a concept representation, that is, the amount of information provided by the representation of a concept with regard to the information defined in the ontology (function Θ).
- The estimator of the commonality of two concept representations (function Ψ).
- The estimator of the difference between two concept representations (function Φ).

In this section, we first detail some guidelines for the practical definition of measures based on the proposed framework. As done in Section 3.3, we define how existing measures can be mapped to the framework and how new proposals can be formulated. Secondly, we use the proposed framework to study and evaluate several SSMs in a biomedical usage context.

4.1. Guidelines for framework instantiation

In this subsection, we define the guidelines to instantiate/design a semantic similarity based on the proposed framework. Two main steps can be distinguished:

1. Selection of an abstract measure, such as $\sum$, σ_α, Sim_{RM} (see Section 3.3, Eqs. (22), (18), and (17)).
2. Definition of the expression of the core elements. This step consists in selecting a specific semantic representation of a concept (ρ function) and the definition of the expression of the abstract operators on which the selected abstract measure relies, for instance to estimate the commonality (Ψ) or the difference (Φ) between two concept representations (e.g. see Table 1).

The first step to design a semantic measure is to select an abstract measure relying on some of the core elements distinguished by framework. The multiple set-based expressions introduced in Section 3.3 and the parameterized abstract measures discussed in Section 3.4 can be used to express a large diversity of measures.

The selected abstract measure defines the semantics related to the compared concepts which will be taken into account during the comparison, e.g. commonalty (Ψ), difference (Φ), and also their weight in the similarity assessment. As an example, both the Jaccard index and the Dice coefficient can be derived from

the Tversky's *ratio model* by setting $\alpha, \beta = 1$ and $\alpha, \beta = 0.5$, respectively. It is therefore explicit that the Dice coefficient gives more importance to commonalities (and less importance to differences) for similarity estimation compared to the Jaccard Index. The selection of the abstract measure is therefore important to finely control the semantics of the scores produced by a measure. This aspect may be particularly important for context-specific applications.

The next step consists in defining how to semantically represent a concept according to the available ontological knowledge. Such representation, i.e. function ρ, is required to define the expression of the operators used by the abstract measure. The selection of a specific representation (ρ function), e.g. set of concepts $A(u)$, partially defines which semantics will be considered in the similarity assessment. Finally, expressions for abstract operators (e.g. estimators of commonalities or differences) must be defined in accordance with the selected expression of function ρ, which defines how to represent a concept.

The users will therefore have to consider (i) specific expressions of the primitive functions distinguished by the framework, (ii) abstract semantic measures (e.g. abstract Tversky's Ratio model) and (iii) specific parameter freedom. Two scenarios can therefore be distinguished:

- The designer has a very clear idea about the more relevant elements that guide the similarity assessment in the concrete scenario and their relative weights and thus tunes and obtains the measure accordingly. Some of the parameters on which the measures rely can, for example, be restricted due to constraints defined by the context of use (e.g. the measure must be symmetric – the user will therefore only consider setting where $\alpha = \beta$ in the abstract ratio model).
- The designer has a training set of similarity scores (human-rated) that would be expected in such a scenario and that can be used to evaluate the accuracy of measures resulting from the framework instantiation. The set of measures to be evaluated can eventually be restricted according to specific properties induced by specific core element expressions, such as the algorithmic complexity (cf. scenario 1). The selection of the best suited measures will therefore be performed empirically using the training set from which performances of measures can be estimated. Such a training set or test sample must be composed of expected scores of similarity for a reasonable amount of pairs of concepts. It must be built alongside the experts of the domain according to the behaviour we want the system to have. In this case, the process detailed in Section 4.2, that is, the calculus of correlation values can be used to select the most appropriate measure. Using the Semantic Measures Library [59] (http://www.semantic-measures-library.org), the users can indeed easily implement a test in which the correlations between various configurations of semantic measures will be evaluated in order to distinguish the most suited set of parameters and therefore the most suited semantic measures according to a specific use case.

With the above-described method, our framework can be used to easily instantiate existing or new SSMs, while finely controlling the semantics considered during the similarity assessment. Such constructive approach draws interesting perspectives for evaluating semantic measures, such as testing the influence of the various SSM components (i.e. abstract measures, core element expressions) over the accuracy of concrete measures in domain-specific tasks.

4.2. Case study in the biomedical domain

As we have seen, once an abstract measure has been selected, e.g. the Tversky's *ratio model*, the various components on which the abstract measure relies need to be defined in order to instantiate a SSM. Among them, we distinguish the parameters of the abstract function, if any (e.g. α, β for the *ratio model*), but also expressions of the core elements (e.g. the representation of a concept ρ the way to assess the commonality of two concept representations Ψ, their difference Φ).

The selection of the parameters governing the measure instantiation is partially driven by the usage context (i.e. ontology). However, values of specific parameters can be difficult to tune and may only be assessed through empirical evaluations. The objective of this section is to discuss the impact of choices made while designing a measure. This study can only be made from the perspective of a specific usage context, since the suitability of SSMs may vary depending on the goal to achieve, in the same manner as human criteria may also vary from one setting to another. As stated in the introduction, due to the importance of SSMs and ontologies in biomedical research, the biomedical domain has been selected as the usage context. Next, we introduce the questions this experiment aims to answer. Then, we detail the experiment design and, finally, we present the results and discussions.

The aim of the experiment is to analyse semantic similarity accuracy in a specific biomedical-related usage context. We specifically want to evaluate which are the parameters of SSMs, distinguished by the framework, which best impacts semantic similarity performance:

- The selection of the abstract measure and its tuning.
- The expression of the core elements of the measures.

This experiment highlights the relevance of the proposed framework to provide answers in a specific usage context. We left additional experiments in other domains, with the use of other measures and parameters for future work.

4.2.1. Experiment design

For this experiment we considered the well-known Pedersen et al. benchmark, which is commonly used in the biomedical domain to evaluate SSMs according to human judgements of similarity [9]. Indeed, to be accurate, most of algorithms and treatments which extensively rely on semantic measures (e.g., in information retrieval, disambiguation and data analysis) require semantic measures to be highly correlated with human judgement of similarity [9,48,49]. Semantic measures are therefore commonly evaluated regarding their ability to mimic human experts' appreciation of similarity between domain-specific concepts. The accuracy of measures is in this case evaluated regarding their correlations to the similarities assessed by domain experts for a set of concept pairs; the more the results of a measure are correlated to the scores of similarity assessed by the experts, the more accurate the measure will be considered.

Pedersen et al. benchmark contains 29 pairs of terms related to the biomedical domain; for each one, the corresponding pair of concepts has been extracted from the SNOMED-CT biomedical ontology [18] (see appendix, Table B1 for the complete list of terms and associated SNOMED-CT concepts). For each pair of concepts, two semantic similarity scores associated to two sets of experts, 9 medical coders and 3 physicians, are given. Those scores have been obtained by averaging the ratings given by the experts of each group. An additional *averaged* score of similarity is also generally considered for each pair, which is computed by averaging the similarities assessed by both coders and physicians. Evaluation of SSMs is then tackled by computing the Pearson correlation against

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

Table 3
Core element expressions evaluated by the experiments.

Instantiation	1	2	3	4
$\rho(u) = \tilde{u}$	G_u^+	$A(u)$	$A(u)$	$A(u)$
$\Theta(\tilde{u})$	$sp(u \rightsquigarrow root)$	$IC(u)$	$\lvert A(u)\rvert$	$\sum_{c \in A(u)} IC(c)$
$\Psi(\tilde{u}, \tilde{v})$	$sp(LCA_{u,v} \rightsquigarrow root)$	$IC(MICA_{u,v})$	$\lvert A(u) \cap A(v)\rvert$	$\sum_{c \in A(u) \cap A(v)} IC(c)$
$\Phi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$

Table 4
Examples of correspondences that can be established between existing SSMs and measure instantiations considering the *contrast/ratio models*, and the instantiation of the core elements defined in Table 3.

Measures	Refs.	Abstract measure	Parameters	Instantiation
Resnik	[14]	Contrast model	$\alpha = 0$, $\beta = 0$	2
Wu and Palmer	[13]	Ratio model	$\alpha = 0.5$, $\beta = 0.5$	1
Lin	[25]	Ratio model	$\alpha = 0.5$, $\beta = 0.5$	2
Sim DIC	[27]	Ratio model	$\alpha = 0.5$, $\beta = 0.5$	4
Pekar and Staab	[20]	Ratio model	$\alpha = 1$, $\beta = 1$	1
Sim Faith	[26]	Ratio model	$\alpha = 1$, $\beta = 1$	2
Concept Match	[24]	Ratio model	$\alpha = 1$, $\beta = 1$	3
Sim_{cGIC}	This paper	Ratio model	$\alpha = 1$, $\beta = 1$	4

Table 5
Best Pearson correlations obtained against coder ratings. Best values obtained using each abstract functions are specified in bold.

Instantiation	Best tuning of *Contrast model*		Best tuning of *Ratio model*		Correlations	
	α	β	α	β	Contrast model	Ratio model
1	0.5	1.0	14.9	2.1	0.764	0.849
2	0.2	0.7	13.6	3.3	**0.801**	0.862
3	0.5	0.4	14.9	3.5	0.613	**0.865**
4	0.4	0.3	8.1	1.9	0.714	0.858

Table 6
Best Pearson correlations obtained against physician ratings. Best values obtained using each abstract functions are specified in bold.

Instantiation	Best tuning of *Contrast model*		Best tuning of *Ratio model*		Correlations	
	α	β	α	β	Contrast model	Ratio model
1	0.2	1.5	6.6	3.2	**0.779**	0.678
2	0.8	0.1	3.6	2.8	0.752	0.683
3	0.3	0.5	3.8	3.4	0.587	0.710
4	0.4	0.4	1.1	1.7	0.670	**0.715**

Table 7
Best Pearson correlations obtained against the average of physician and coder ratings.Best values obtained using each abstract functions are specified in bold.

Instantiation	Best tuning of *Contrast model*		Best tuning of *Ratio model*		Correlations	
	α	β	α	β	Contrast model	Ratio model
1	0.3	1.3	14.9	2.6	0.799	0.789
2	0.5	0.4	6.9	3.2	**0.805**	0.798
3	0.4	0.4	7.9	3.7	0.623	**0.810**
4	0.4	0.4	2.8	2.0	0.719	0.808

the similarity ratings given by each group of human experts (physicians, coder, both). In this experiment, the semantic similarity of each pair of concepts has been computed using SNOMED-CT as ontology.

The study focuses on two abstract measures: the *contrast model* (Sim_{CM}) and the *ratio model* (Sim_{RM}) (equations 16 and 17).The α and β parameters, which tune the contribution of the information found in u (resp. v) which is not found in v (resp. u), were set from 0 to 15 with a step of 0.1, i.e., 150 α and β values. As a result of this parameter tuning, 22,500 abstract expressions ($150 * 150$) of both Sim_{CM} and Sim_{RM} have been obtained and systematically evaluated. Notice that in Sim_{CM}, the γ parameter, which tunes the contribution of the commonality, was fixed to 1 as we did not want to study the effect of the variation of both the commonality and the difference (α and β) since they are both inversely correlated in most cases. For each abstract expression among the 22,500 evaluated, we further tested the four instantiations of the core elements shown in Table 3.

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

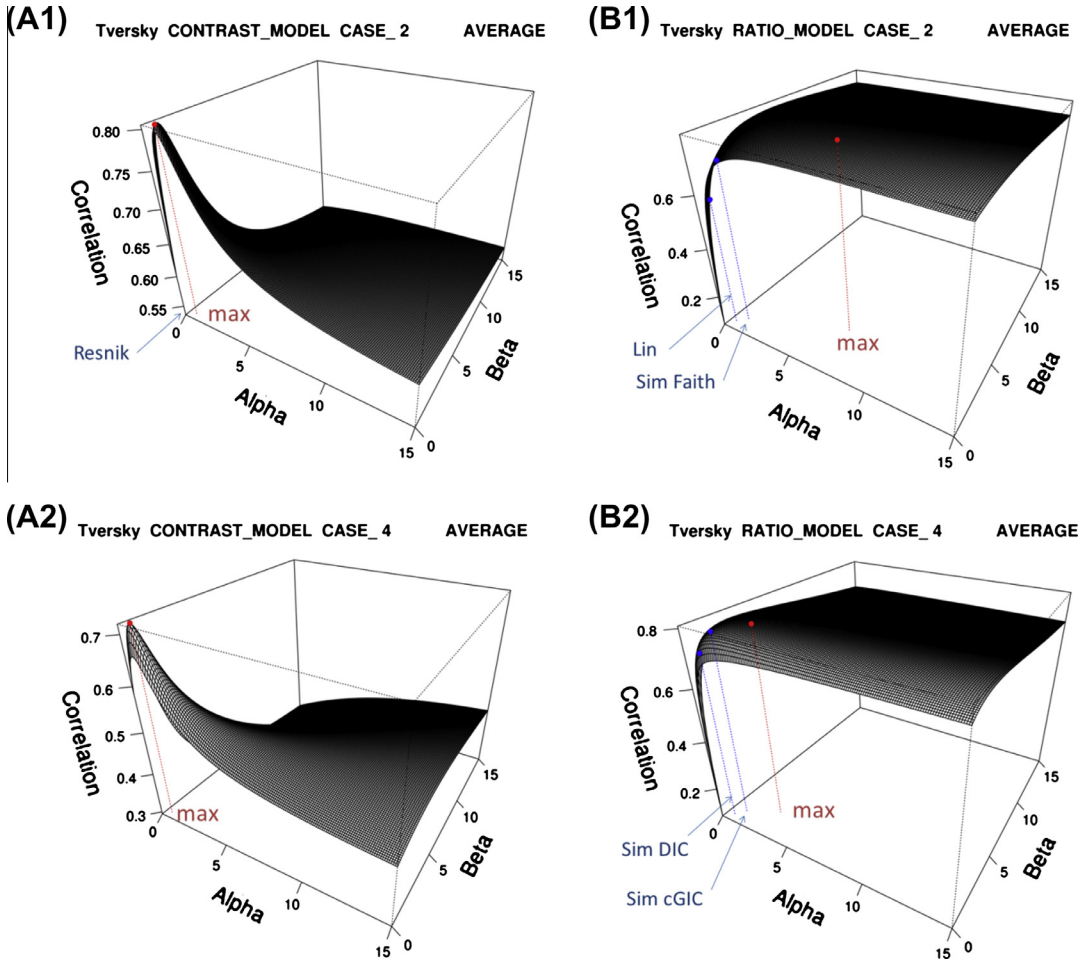


Fig. 5. Surfaces associated to the Pearson correlations of similarity measures against the average of physician and coder ratings. Measures have been instantiated from the *contrast model* (A) and the *ratio model* (B) using core elements expressions defined Table 3: expression #2 (A1, B1) and expression #4 (A2, B2). Each point composing the surface corresponds to a specific tuning of α and β. For each surface, the dot labeled *max* corresponds to the maximal value observed. Other dots reflect instantiations that correspond to existing measures, cf. labels and Table 4.

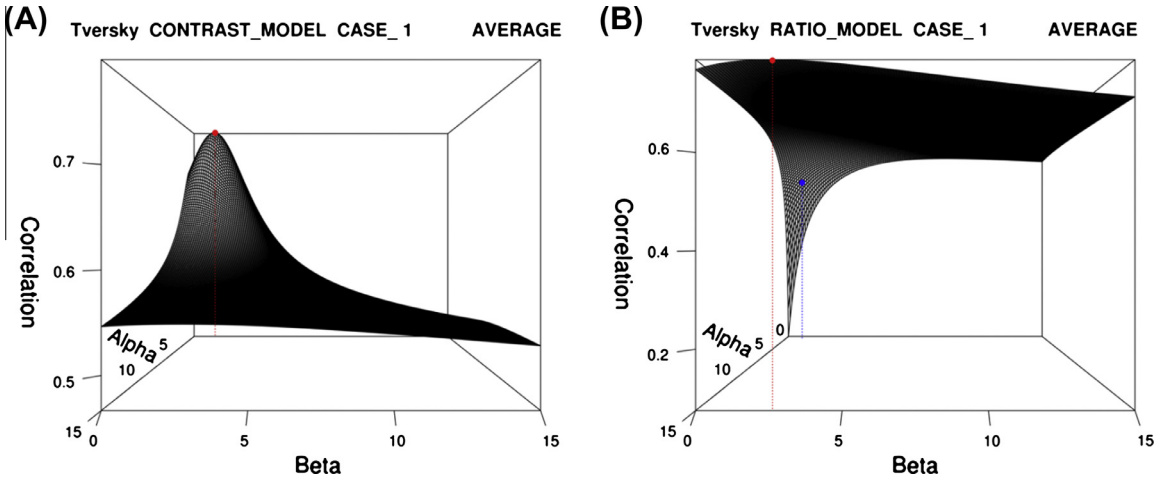


Fig. 6. Plot of the Pearson correlations of the *contrast* and the *ratio model* using instantiation #1 (Averaged benchmark).

Thus, for each abstract similarity measure, the four instantiations of the core elements lead to 90,000 individual measures, i.e. 22,500 * 4. Note that IC-dependent configurations used Sánchez et al. IC calculus model [31]. The final experiment is then based on the evaluation of more than 1 million measure configurations, i.e. 360,000 measure configurations for each evaluation benchmark: physicians, coders and the average of both ratings.

Notice that some measures available in the literature correspond to particular points into the range of measure instantiations performed in this experiment. Table 4 highlights some of these correspondences.

Empirical evaluations were performed using the Semantic Measures Library [59][1] a library dedicated to large-scale analysis and computation of SSMs. The source code and detailed documentation related to the experiment is open sourced and available at http://www.lgi2p.ema.fr:8090/~sharispe/publications/JBI2013.

4.2.2. Results and discussions

Tables 5–7 summarize the best results that have been obtained for each configuration of abstract measures and for the four specific strategies used to express the core elements (cf. Table 3). The correlations have been computed against averaged scores obtained considering only coders scores (Table 5), only physicians scores (Table 6) and the average of both coders and physicians scores (Table 7).

Some conclusions can be extracted from the analysis of the results:

- *The effect of core elements' expressions on measures' accuracies depend on the abstract measure considered*: for the *contrast model*, the core elements' expressions corresponding to instantiation #3 always resulted in the lowest correlations (0.613, 0.587, 0.623). On the contrary, considering the *ratio model*, instantiation #3 resulted in some of the best correlations (0.865, 0.710, 0.810).
- *The accuracy of the measures is mainly explained by the selected abstract measure*: indeed, changes in the expression of the core elements only modified the shape of the surface slightly. Moreover, most instantiations associated with well-tuned α and β parameters produced good correlations. The maximum variation between the best correlation observed for the *ratio model* was ±0.04 (0.678–0.715, see Table 6). However, using the *contrast model*, greater variations were observed: ±0.19 (0.587–0.779, see Table 6). We can therefore conclude that relevancy of tuning α and β depends on the selected abstract measure from which the similarity measure will be instantiated.
- *The variability of scores is mainly due to the selection and tuning of the abstract measure*: by considering the *contrast model*, an important variability of results is observed depending on the values of α and β (see Fig. 5A1 and A2). However, despite the variability of the results, it is also observed for the *ratio model* that the variability significantly decreases with large values of α and β (see Fig. 5B1 and B2). It is therefore interesting to remark that the sensibility of measures does not depend on the expression of the abstract operators, but on the selected abstract measure.
- *Asymmetrical measures tend to provide the best results*: all experiments provided the best correlations by tuning the measures with asymmetric contributions of α and β parameters (see Tables 5–7 and Fig. 6). In Fig. 6, the asymmetry of the surfaces underlines the benefits of considering asymmetric α and β values. The improvement of an asymmetric tuning of parameters is best outlined in the *ratio model* (Fig. 6 B). Such result stresses the need of tools to best estimate parameters for domain-specific settings.

Note that the observations made in this experiment are only driven by the analysis of specific configurations of measures, using a single ontology and a unique benchmark. However, our results stress the usefulness of such experiments and the added value of the proposed framework to both ease and improve SSMs understanding, selection and design. Such tool is also essential to optimize SSMs for domain-specific usage contexts and to facilitate selection of specific measures.

5. Conclusions and future work

A large diversity of semantic similarity measures (SSMs) has been proposed over the last decades, most of them focused on specific applications or domains. In this paper, we unified most well-known approaches through the definition of a theoretical framework dedicated to SSMs. The main advantages of the proposed framework rely on *the identification of the core elements commonly used to design SSMs*. We have indeed underlined that *most measures can be expressed considering a limited set of core elements (functions)* such as those defining (i) how to represent a concept through a processable canonical form (ρ), (ii) how to estimate its specificity (θ) and the specificity of its representation (Θ), and (iii) how to estimate the degree of commonality (Ψ) and difference (Φ) between two concept representations. In fact, we demonstrate how those core elements can be used to express a large diversity of (existing) measures based on generic parametric measures which can be seen as the backbone of semantic measures. The characterization of the measures through the distinguished core elements enabled us to better characterize measures relying on different paradigms and therefore to better understand the large diversity of measure expressions proposed in the state-of-the-art. In addition to this contribution, we also bring out, through detailed case studies and examples, the practical applications and interesting perspectives this framework provides:

- *Theoretical analysis and understanding of semantic measures.* Distinguishing the core elements on which SSMs are based allow us to highlight narrow relationships between existing proposals. Indeed, we found that SSMs can be easily expressed through the definition of a few intuitive core elements and that most, if not all, measures are just particular expressions of a limited set of abstract measures. We therefore demonstrated that several measures which rely on the same abstract measure (e.g., abstracted ratio model), only differ due to a specific set of parameters selected to instantiate them (e.g., strategy used to represent a concept or to assess the commonality/difference between concept

Table A1
Selection of SSMs adopting the edge-counting approach. IOI = Identity of the Indiscernibles, the * symbol denotes that the measure is a distance.

Name	Refs.	Range	IOI
Rada et al.*	[15]	$\mathbb{R}^+$	True
Pekar and Staab	[20]	[0, 1]	True
Wu and Palmer	[13]	[0, 1]	True
Slimani et al.	[50]	[0, 1]	True
Nagar and Al-mubaid	[51]	$\mathbb{R}^+$	True

Table A2
Selection of SSMs adopting the feature-based strategy. IOI = Identity of the Indiscernibles, the * symbol denotes that the measure is a distance.

Name	Refs.	Range	IOI
Sanchez et al.*	[22]	[0, 1]	True
Ranwez et al.*	[52]	$\mathbb{R}^+$	True
Batet et al.	[53]	[0, 1]	True
Rodriguez and Egenhofer	[23]	[0,1]	True
Petraskis et al.	[54]	[0,1]	True
Maedche and Staab	[24]	[0,1]	True

[1] See dedicated website http://www.semantic-measures-library.org.

Table A3
Selection of SSMs based on Information Theory. IOI = Identity of the Indiscernibles, the * symbol denotes that the measure is a distance; *1: the range of Resnik similarity measures depends on the selected IC function.

Name	Refs.	Range	IOI
Resnik	[14]	$\mathbb{R}^{+*1}$	False
Jiang and Conrath*	[12]	[0, 1]	True
Lin	[25]	[0, 1]	True
Sim Rel	[55]	[0, 1]	False
Sim DIC	[27]	[0, 1]	True

Table A4
Selection of SSMs based on the hybrid approaches. IOI = Identity of the Indiscernibles.

Name	Refs.	Range	IOI
Li et al.	[37]	$\mathbb{R}^{+}$	False
Mao and Chu	[56]	$\mathbb{R}^{+}$	False
SSM	[36]	[0, 1]	False
Al-Mubaid and Nguyen	[57]	$\mathbb{R}^{+}$	True
SSA	[58]	[0, 1]	False

representations). This strong result is therefore important for the theoretical analysis of semantic measures. Indeed, most applications in which the measures are not selected through empirical analyses, expect the measures to fulfil specific properties, e.g., symmetry, respect of the identity of the indiscernible. Thanks to the breakdown of the measures proposed by the framework, properties of measures can be analysed, not only regarding specific measure instantiation (e.g. Lin's, Resnik's), but also focusing on both the abstract measures from which they derived (e.g. abstract contrast model) and the properties induced by the core elements. This enables a better understanding and analysis of the algorithmic complexity of measures, which is critical for most application contexts but nevertheless rarely considered in semantic measure proposals.

- *Creation and tuning of semantic measures*. The separation of measures from the core elements on which they rely enables researchers to focus not just on new ad hoc measures, but also on the design of specific strategies to improve the assessment of those core elements. As an example, we have seen that an accurate estimator of the commonality between two concepts (Ψ function), which depends on the canonical form adopted to represent a concept (ρ function), is of major importance to define semantic measures. Designers of measures can therefore improve several existing measures by improving the way the Ψ function is estimated with regard to a specific representation of a concept. It is therefore important to understand that improving the assessment of core elements distinguished by the framework leads to improvements in *multiple* measures, and not just to a specific measure in a concrete context. By distinguishing the core elements of semantic measures, the theoretical tool proposed in the paper therefore opens interesting perspectives for the definition and improvement of semantic measures in general. Moreover, by comparing measure performances in a biomedical-related context, we also illustrate how the framework can be used to express parameterized measures and to guide the adoption of a specific strategy. For example, if one wants to design an application aiming to cluster documents according to the similarity of their semantic annotations, he will have to select an appropriate measure for this task. With the help of our framework, instead of asking experts to evaluate the clusters produced by various similarity measures, which is time consuming, one can just

Table B1
Pedersen et al. benchmark used to evaluate semantic measures using SNOMED-CT ontology (version 2013). The SNOMED-CT concept identifiers corresponding to each terms involved in the benchmark are given. 29 pairs have been used, the average similarity associated to each pairs by the groups of Physicians (Phy.), Coders (Cod.) and Physicians + Coders (Avg.) are also reported.

Term 1	Term 2	Concept 1	Concept 2	Phy.	Cod.	Avg.
Renal failure	Kidney failure	42399005	42399005	4	4	4
Heart	Myocardium	80891009	74281007	3.3	3	3.15
Stroke	Infarct	230690007	55641003	3	2.8	2.9
Abortion	Miscarriage	70317007	17369002	3	3.3	3.15
Delusion	Schizophrenia	48500005	58214004	3	2.2	2.6
Congestive heart failure	Pulmonary edema	42343007	19242006	3	1.4	2.2
Metastasis	Adenocarcinoma	128462008	443961001	2.7	1.8	2.25
Calcification	Stenosis	125369001	415582006	2.7	2	2.35
Diarrhea	Stomach cramps			2.3	1.3	1.8
Mitral stenosis	Atrial fibrillation	79619009	49436004	2.3	1.3	1.8
Chronic obstructive pulmonary disease	Lung infiltrates			2.3	1.9	2.1
Rheumatoid arthritis	Lupus	69896004	200936003	2	1.1	1.55
Brain tumor	Intracranial hemorrhages	254935002	1386000	2	1.3	1.65
Carpel tunnel syndrome	Osteoarthritis	57406009	396275006	2	1.1	1.55
Diabetes mellitus	Hypertension	73211009	38341003	2	1	1.5
Acne	Syringes	11381005	61968008	2	1	1.5
Antibiotic	Allergy	255631004	106190000	1.7	1.2	1.45
Cortisone	Total knee replacement	32498003	179344006	1.7	1	1.35
Pulmonary fibrosis	Lung cancer	51615001	363358000	1.7	1.4	1.55
Cholangiocarcinoma	Colonoscopy	70179006	73761001	1.3	1	1.15
Lymphoid hyperplasia	Laryngeal Cancer	128863005	363429002	1.3	1	1.15
Multiple sclerosis	Psychosis	24700007	69322001	1	1	1
Appendicitis	Osteoporosis	74400008	64859006	1	1	1
Rectal polyp	Aorta			1	1	1
Xerostomia	Alcoholic cirrhosis	87715008	420054005	1	1	1
Peptic ulcer disease	Myopia	13200003	57190000	1	1	1
Depression	Cellulites	35489007	128045006	1	1	1
Varicose vein	Entire knee meniscus			1	1	1
Hyperlipidemia	Metastasis	55822004	363346000	1	1	1

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006

ask them to rate the similarity of pairs of individual concepts and use them as training data to systematically evaluate and select appropriate measures, as done in our experiments. Therefore, the proposed approach can be used to select the most appropriate measures according to particular criteria by driving the decision process, much in the spirit of learning theories (e.g. correlation with respect to expected values given by an expert, algorithmic complexity of measures, etc.).

We are currently investigating semantic similarity performances through the new insight provided by the proposed framework. Considering the large diversity of measures available, an important contribution for end-users of SSMs would be to provide tools which enable to select the best-suited measures for domain-specific usage contexts. The framework presented in this paper provides the theoretical frame for developing such tool. In addition, the framework will be used to perform detailed evaluations in other contexts and applications (i.e. other knowledge domains, ontologies and training data). We also plan to extend the framework to support semantic measures other than SSMs, i.e. relatedness [9], by exploiting non-taxonomic relationships available in the ontology and to compare groups of concepts, rather than pairs, which are widely used, for instance, to compare genes annotations [7].

Acknowledgments

This work was partly supported by the French Life Sciences and Healthcare Alliance (AVIESAN), the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects ICWT TIN2012-32757, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia (under Grant 2009 SGR 1135).

Appendix A

See Tables A1–A4.
See Table B1.

References

[1] Gruber T. A translation approach to portable ontology specifications. Knowledge Acquisition 1993;5(2):199–220.
[2] Sy M-F, Ranwez S, Montmain J, Regnault A, Crampes M, Ranwez V. User centered and ontology based information retrieval system for life sciences. BMC Bioinformatics 2012;13(Suppl 1):S4.
[3] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EGM, Milios E. Information retrieval by semantic similarity. Int J Semant Web Inf Syst 2006;2:55–73.
[4] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proc Fourth Int Conf Intell Text Process Comput; 2003. p. 241–57.
[5] Leroy G, Rindflesch TC. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. Int J Med Inform 2005;74:573–85.
[6] Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 2011;7:496.
[7] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009;5:12.
[8] Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J Biomed Inform 2011;44:749–59.
[9] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40:288–99.
[10] Bizer C, Heath T, Berners-Lee T. Linked data – the story so far. Int J Semant Web Inf Syst 2009;5:1–22.
[11] Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinform 2012;13:569–85.
[12] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Int Conf Res Comput Linguist (ROCLING X); 1997. p. 15.
[13] Wu Z, Palmer M. Verb semantics and lexical selection. In: 32nd. Annu Meet Assoc Comput Linguist; 1994. p. 133–8.
[14] Resnik P. Using Information content to evaluate semantic similarity in a taxonomy. In: Proc 14th Int Jt Conf Artif Intell IJCAI; 1995. p. 448–53.
[15] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;19:17–30.
[16] Blanchard E, Harzallah M. A typology of ontology-based semantic measures, EMOI-INTEROP'05. In: Proc Open Interop Work Enterp Model Ontol Interoperability; 2005.
[17] Blanchard E, Harzallah M, Kuntz P. A generic framework for comparing semantic similarities on a subsumption hierarchy. 18th Eur Conf Artif Intell; 2008. p. 20–4.
[18] Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthc Informatics Bus Mag Inf Commun Syst 2004;21:54–6.
[19] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: Fellbaum C, editor. WordNet an electron, Lex database. MIT Press; 1998. p. 265–83.
[20] Pekar V, Staab S. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: COLING '02 Proc 19th Int Conf Comput Linguist Association for, Computational Linguistics; 2002. p. 1–7.
[21] Tversky A. Features of similarity. Psychol Rev 1977;84:327–52.
[22] Sánchez D, Batet M, Isern D, Valls A. Ontology-based semantic similarity: a new feature-based approach. Expert Syst Appl 2012;39:7718–28.
[23] Rodríguez A, Egenhofer MJ. Determining semantic similarity among entity classes from different ontologies. IEEE Trans Knowl Data Eng 2003;15:442–56.
[24] Maedche A, Staab S. Comparing ontologies – similarity measures and a comparison study (internal report). Karlsruhe; 2001.
[25] Lin D. An information-theoretic definition of similarity. In: 15th Int Conf Mach Learn, Madison, WI; 1998. p. 296–304.
[26] Pirró G, Euzenat J. A feature and information theoretic framework for semantic similarity and relatedness. In: Proc 9th Int Semant Web Conf ISWC 2010, Springer; 2010. p. 615–30.
[27] Mazandu Gaston K, Mulder Nicola J. IT-GOM: an integrative tool for IC-based GO semantic similarity measures; 2011.
[28] Lee JH, Kim MH, Lee YJ. Information retrieval based on conceptual distance in is-a hierarchies. J Doc 1993;49:188–207.
[29] Couto FM, Silva MJ, Coutinho PM. Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In: Conf Inf Knowl Manag, ACM; 2005. p. 343–4.
[30] Couto FM, Silva MJ. Disjunctive shared information between ontology concepts: application to gene ontology. J Biomed Semantics 2011;2:5.
[31] Sánchez D, Batet M, Isern D. Ontology-based information content computation. Knowledge-Based Syst 2011;24:297–303.
[32] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet. In: 16th Eur Conf Artif Intell. IOS Press; 2004. p. 1–5.
[33] Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet. In: FGCNS '08 Proc 2008 Second Int Conf Futur Gener Commun Netw Symp – vol. 03, IEEE Computer Society; 2008. p. 85–9.
[34] Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274–81.
[35] Couto FM, Silva M, Coutinho PM. Implementation of a functional semantic similarity measure between gene-products. Department of Informatics, University of Lisbon; 2003.
[36] Othman RM, Deris S, Illias RM. A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. J Biomed Inform 2008;41:65–81.
[37] Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 2003;15:871–82.
[38] Choi S, Cha S, Tappert CC. A survey of binary similarity and distance measures. J Syst Cybern Informatics 2010:43–8.
[39] Cross V, Yu X. A fuzzy set framework for ontological similarity measures. Comput Intell 2010:18–23.
[40] Cross V, Yu X, Hu X. Unifying ontological similarity measures: a theoretical and empirical investigation. Int J Approx Reason 2013;54:861–75.
[41] Cross V. Tversky's parameterized similarity ratio model: a basis for semantic relatedness. In: Fuzzy Inf Process Soc 2006. NAFIPS 2006. Annu Meet. North Am, Montreal, Que; 2006. p. 541–6.
[42] Cross V. Fuzzy semantic distance measures between ontological concepts. In: IEEE Annu Meet Fuzzy Information, 2004. Process NAFIPS '04, IEEE, vol. 2; 2004. p. 635–40.
[43] Pesquita C, Faria D, Bastos H. Evaluating gobased semantic similarity measures. In: Proc 10th Annu Bio- 2007; 2007. p. 37–40.
[44] Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 1999;11:95–130.
[45] Yang H, Nepusz T, Paccanaro A. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics 2012;28:1383–9.

[46] Lin D. Automatic retrieval and clustering of similar words. In: Proc 17th Int Conf Comput Linguist – vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA; 1998. p. 768–74.
[47] Webster RJ. Convexity; 1994.
[48] Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. AMIA Annu Symp Proc 2010;2010:572–6.
[49] Pakhomov SVS, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 2011;44:251–65.
[50] Slimani T, Boutheina BY, Mellouli K. A new similarity measure based on edge counting. World Acad Sci Eng Technol 2006:34–8.
[51] Nagar A, Al-Mubaid H. A new path length measure based on GO for gene similarity with evaluation using SGD pathways. In: 2008 21st IEEE Int Symp Comput Med Syst, IEEE; 2008. p. 590–5.
[52] Ranwez S, Ranwez V, Villerd J, Crampes M. Ontological distance measures for information visualisation on conceptual maps. Lect Notes Comput Sci 2006;4278(2006):1050–61.
[53] Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2010;4:39–52.
[54] Petrakis E, Varelas G, Hliaoutakis A, Raftopoulou P. X-similarity: computing semantic similarity between concepts from different ontologies. J Digit Inf Manage 2006;4:233.
[55] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 2006;7:302.
[56] Mao W, Chu WW. Free-text medical document retrieval via phrase-based vector space model. In: AMIA Symp Am Med Informatics Assoc; 2002. p. 489–93.
[57] Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. In: Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf; 2006. p. 2713–7.
[58] Alvarez M, Yan C. A graph-based semantic similarity measure for the gene ontology. J Bioinform Comput Biol 2011;9:681–95.
[59] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies, Bioinformatics, 2013.

Please cite this article in press as: Harispe S et al. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J Biomed Inform (2013), http://dx.doi.org/10.1016/j.jbi.2013.11.006