Towards the estimation of feature-based semantic similarity using multiple ontologies

Albert Solé-Ribalta[1], David Sánchez, Montserrat Batet, Francesc Serratosa

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26. 43007 Tarragona (Spain)

Abstract

A key application of ontologies is the estimation of the semantic similarity between terms. By means of this assessment, the comprehension and management of textual resources can be improved. However, most ontology-based similarity measures only support a single input ontology. If any of the compared terms do not belong to that ontology, their similarity cannot be assessed. To solve this problem, multiple ontologies can be considered. Even though there are methods that enable the multi-ontology similarity assessment by means of integrating concepts from different ontologies, most of them are based on simple terminological and/or partial matchings. This hampers similarity measures that exploit a broad set of taxonomic evidences of similarity, like feature-based ones. In this paper, we tackle this problem by proposing a method to identify *all* the suitable matchings between concepts of different ontologies that intervene in the similarity assessment. In addition to the obvious terminological matching, we exploit the ontological structure and the notion of concept subsumption to discover non-trivial equivalences between heterogeneous ontologies. Our final goal is to enable the accurate application of feature-based similarity measures in a multi-ontology setting. Our proposal is evaluated with regard human judgements of similarity for several benchmarks and ontologies. Results shows an improvement against related works, with similarity accuracies that even rival those obtained in an ideal mono-ontology setting.

Keywords: feature-based semantic similarity, multiple ontologies, WordNet, MeSH

[1] Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain. Tel.: +34 977 558676;
E-mail: albert.sole@urv.cat.

1. Introduction

General-purpose ontologies such as WordNet (Fellbaum, 1998), or domain-specific ones like MeSH (Nelson, et al., 2001), provide a formal and machine readable representation of knowledge that can be used in a variety of tasks in which a semantic interpretation of text is required, such as information extraction (Sánchez, et al., 2011) and retrieval (Pedersen, et al., 2007), semantic data mining (Batet, 2011) or privacy-preserving methods (Martínez, et al., 2012; Sánchez, et al., 2013). The main use of ontologies, on which most of the above tasks rely, is the computation of the *semantic similarity* between textual terms. Semantic similarity measures quantify the degree of taxonomic resemblance between a pair of terms (e.g. *flu* and *bronchitis* are similar because both are *disorders of the respiratory system*), by analysing taxonomic relationships modelled in an ontology.

Throughout the years, many ontology-based similarity measures have been developed. These can be classified into different families according to their theoretical principles. *Edge-counting measures* quantify similarity according to the *length of the shortest path* defined by the taxonomic relationships that separate two concepts in an ontology (Rada, et al., 1989; Wu & Palmer, 1994), as exemplified in Figure 1. Since these measures do not evaluate paths other than the shortest one, their accuracy is limited (Sánchez, et al., 2012a). To benefit from additional knowledge that ontologies provide, *feature-based measures* quantify similarity according to the amount of common and/or non-common taxonomic subsumers of the compared concepts (Maedche & Zacharias, 2002; Sánchez, et al., 2012a), as shown in Figure 2. Thanks to the evaluation of additional taxonomic knowledge than edge-counting measures, feature-based measures achieve a higher accuracy (Sánchez, et al., 2012a).

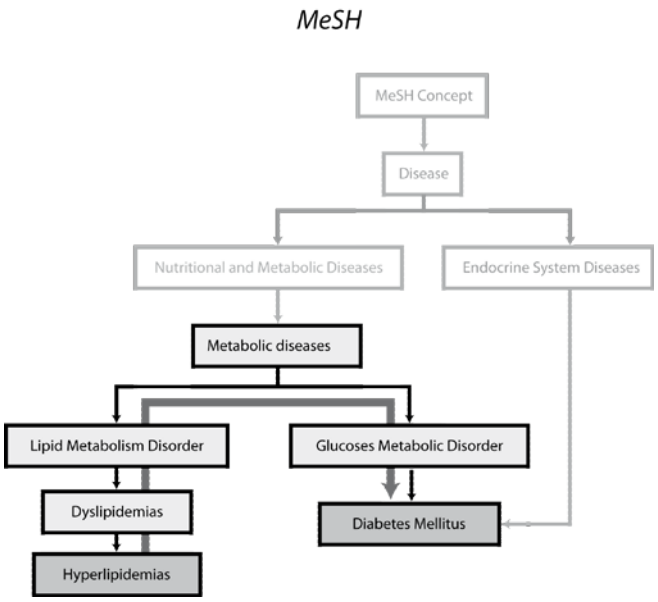

Figure 1. Similarity assessment for *Hyperlipidemias* and *Diabetes Mellitus* in MeSH ontology with an *edge-counting* similarity measure. The thick arrow shows the taxonomic path evaluated in the similarity assessment.

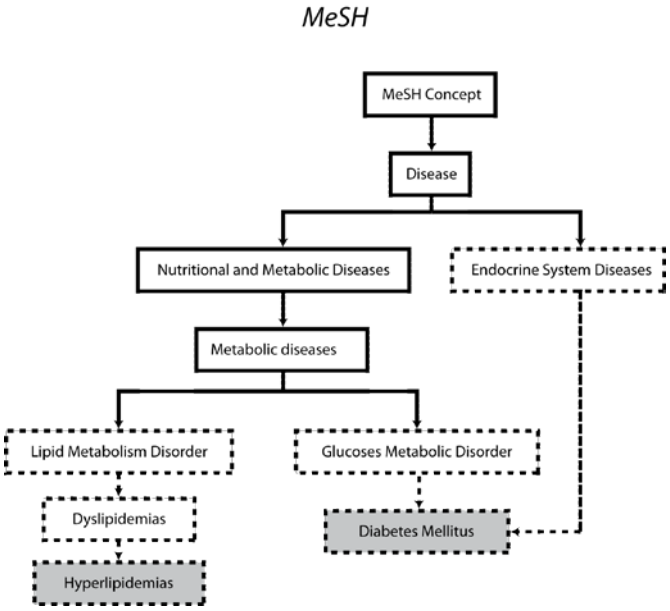

Figure 2. Similarity assessment for *Hyperlipidemias* and *Diabetes Mellitus* in MeSH ontology with a *feature-based* similarity measure. Concepts within dashed squares correspond to taxonomic subsumers that the compared concepts do not have in common, whereas concepts within solid squares represent common subsumers.

However, these measures are limited by the coverage and detail of the input ontology (Sánchez, et al., 2012b). If any of the compared terms cannot be found in the input ontology, the similarity

cannot be assessed. As acknowledged by several authors (Al-Mubaid & Nguyen, 2009; Batet, et al., 2013; Petrakis, et al., 2006; Sánchez, et al., 2012b), this limitation can be overcome by exploiting *multiple ontologies*. In this scenario, each term belongs to a different ontology. Thus, the basic idea is to discover *common taxonomic subsumers between ontologies* to later apply similarity measures like in a mono-ontology setting (Rodríguez & Egenhofer, 2003). Most works framed in the multi-ontology scenario rely on terminological matching to discover common taxonomic subsumers (Al-Mubaid & Nguyen, 2009; Batet, et al., 2013; Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003). However, since ontologies rarely model concepts in the same way, or refer to them by using the same label (due to synonymy), strategies based on terminological matching omit part of the equivalent concepts, a circumstance that tends to cause an underestimation of similarity. Moreover, most authors solely focus on discovering a *unique* common subsumer (i.e. the least common one) (Al-Mubaid & Nguyen, 2009; Sánchez, et al., 2012b), so that the shortest path between concepts can be computed to apply edge-counting measures. This fact hampers feature-based measures because they rely on additional taxonomic evidences whose correspondences across the different ontologies will remain unknown.

To tackle these problems, in this paper we propose a method for enabling similarity assessments from multiple ontologies, which offers the following contributions:

- It aims to identify *all* the suitable correspondences between taxonomic subsumers of a pair of concepts belonging to different ontologies (and not just a unique subsumer, as it is done in most related works). In this manner, feature-based measures can be accurately applied in a multi-ontology setting.
- It exploits the structural and semantic resemblances between ontologies to discover equivalences between subsumers that are conceptually equivalent and not just terminologically identical.
- Since scopes, sizes and detail levels may vary from one ontology to another, our method supports the integration of taxonomic structures with different granularity degrees; that is, it is able to find, if appropriate, multiple correspondences of a single subsumer.
- It incorporates an iterative algorithm that is able to dynamically adapt the matching process according to already discovered correspondences between subsumers.

The evaluation, which is based on applying several feature-based measures in a multi-ontology scenario and on comparing the obtained results with human judgements of similarity, shows an improvement of similarity accuracy against related works.

The rest of the paper is organised as follows. Section 2 reviews related works aimed to enable the similarity assessment (feature-based and also edge-counting methods) across multiple ontologies. Section 3 analyses their main shortcomings and presents the hypothesis upon our method relies. Section 4 formalises the proposed method. Section 5 presents the evaluation, which considers several benchmarks of biomedical terms and compares and discusses the obtained results against related works. The final section contains the conclusions and proposes some lines of future research.

2. Related work

The first works to deal with multiple ontologies were based on either mapping terms of different ontologies into a single ontology (Gangemi, et al., 1998; Weinstein & Birmingham, 1999) or on creating a new ontology by merging existing ones (Bergamaschi, et al., 1998; Gangemi, et al., 1998; Mena, et al., 1996). However, complete ontology integration represents a challenging problem from the cost and scalability points of view, due to large ontologies contain thousands of concepts (Fellbaum, 1998; Nelson, et al., 2001), and due to the difficulty to deal with overlapping concepts and inconsistencies when dealing with the complete ontological structure (Rodríguez & Egenhofer, 2003).

If we tackle the problem from a more convenient perspective, we will notice that some authors have focused solely on finding correspondences between those parts of the ontologies that affect the similarity assessment, that is, the taxonomic trees to which compared concepts belong. The most basic approaches create an imaginary root node that subsumes all the concepts of the different ontologies (Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003). In (Rodríguez & Egenhofer, 2003), authors evaluate features (synonyms, meronyms and attributes) of concepts and those of their neighbourhood. Overlapping features are detected by means of terminological matching of terms of different ontologies, and term resemblance is computed as the weighted sum of their degree of feature overlapping. Since non-taxonomic features are considered in the assessment, in that work, *semantic relatedness* (instead of *similarity*) is quantified. Petrakis et al.

(Petrakis, et al., 2006) extended the previous approach by relying on terminological matching between synonym sets and concept glosses (i.e. words extracted by parsing term definitions). Term relatedness was computed with the Jaccard coefficient.

Authors in (Batet, et al., 2013) also rely on terminological matching to discover the most specific pair of equivalent subsumers. Then, those subsumers that are more general than the equivalent pair are also assumed to be equivalent (because they generalise the equivalent pair), whereas those subsumers that are more specific are considered different. This strategy considers that the subsumption relation should be propagated to concept ancestors. That is, if a concept pair is equivalent, their ancestors should also be equivalent. Even though this strategy overcomes some of the limitations of the strict terminological matching of ancestors, it could fail to detect subtle semantic differences. General subsumers are considered systematically as equivalent, which could not be the case, especially when modelling multiple inheritances.

Approaches relying on edge-counting measures focus on connecting taxonomic trees of different ontologies by searching out a pair of concepts subsuming the compared ones that can act as a bridge between ontologies. As a result, edge-counting measures, which measure the length of the shortest taxonomic path, can be directly applied, as in (Al-Mubaid & Nguyen, 2009). In that work, authors connect ontologies by joining taxonomic subsumers with the same textual label. Then, the maximum similarity obtained from the paths defined by the set of bridges is taken. Because the path length provides absolute values, the method relies on the selection of a predefined primary ontology (the rest are considered secondary) to which similarity values are normalised. Authors scale the part of the path computed from the secondary ontology to the primary ontology. The scaling factor is the difference of the taxonomic depth of the secondary ontology versus the primary one. A disadvantage of this methodology is the fact that a primary ontology must be selected a priori by the user. A similar method is proposed in (Saruladha, et al., 2010). These authors consider both ontologies equally important and analyse the set of subsumers of the compared concepts to discover terminological matchings.

In (Sánchez, et al., 2012b), authors also propose a method to enable the application of edge-counting measures in a multi-ontology setting. To do so, the authors discover the most similar subsumer pair that acts as *the least common subsumer* (LCS) for path length calculus. Contrary to the above methods, authors do not solely rely on terminological matching. Subsumer pairs are

compared according to their degree of semantic overlapping and structural similarity. Regarding the former, authors argue that, since the set of hyponyms of a concept summarises and binds its meaning, thus allowing to differentiate it from other ones (Sánchez & Batet, 2012) (Blank, 2003), a subsumer pair sharing an amount of hyponyms gives evidences of resemblance. Structural similarity between subsumer pairs is evaluated as the aggregation of the degree of semantic overlapping of their adjacent concepts. The hypothesis is that, if two subsumers are similar, their adjacent concepts (i.e. direct hyponyms and hypernyms) should also be similar. The pair of subsumers that maximise the average of both features is taken as the equivalent least common subsumer (LCS) between the two ontologies. Thanks to the less constrained subsumer matching, this last approach enables more accurate similarity estimations than the above-discussed methods.

3. Problem analysis

As discussed above, most related works rely on the terminological matching of concept labels in order to discover common subsumers between different ontologies, and/or solely focus on the detection of a unique subsumer pair that could act as a bridge between ontologies. On the one hand, approaches relying on terminological matching tend to underestimate the real similarity between concepts. Due to language ambiguity, conceptually identical terms may be referred by different textual labels (e.g. synonyms, acronyms, lexicalisations, etc.). Moreover, since ontologies rarely model knowledge in the same way, subsumers do not completely overlap and, hence, they may be referred with non-perfectly synonymous labels (e.g. *cancer* and *tumour*, although very similar, are not completely identical because *cancer* is considered a malignancy whereas a *tumour* may be benign). As a consequence, potentially equivalent concepts may be omitted and, thus, commonalties between concepts may not be properly evaluated. Figure 3 illustrates this problem: the first pair of terminologically identical subsumers for *antibiotic* (in WordNet) and *anti-bacterial agent* (in MeSH) is *drug*. However, their similarity is underestimated because the pair *bactericide* (in WordNet) and *anti-bacterial agent* (in MeSH) are almost synonymous. Thus, for edge-counting approaches, concepts may seem more distant because the most appropriate least common subsumer could not be detected (Sánchez, et al., 2012b).

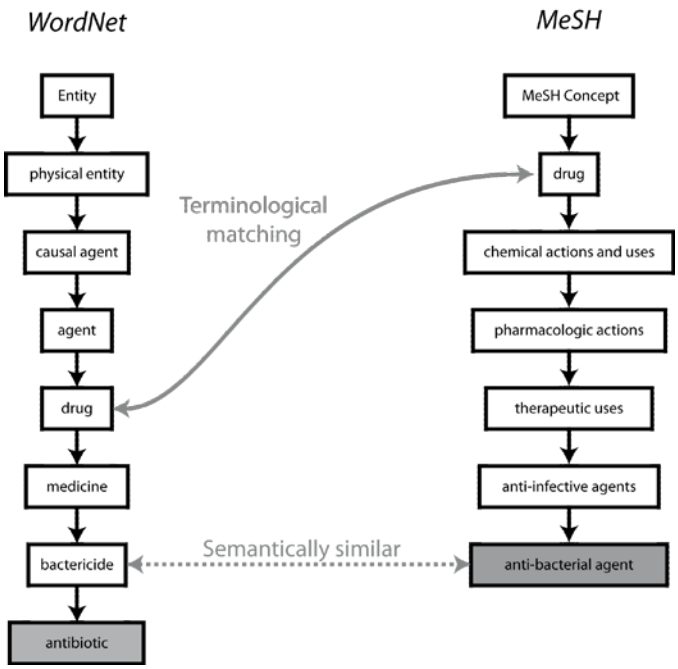

Figure 3. Example of terminological matching between subsumers of *antibiotic* (in WordNet) and *antibacterial agent* (in MeSH).

On the other hand, the fact that a unique subsumer pair is detected, like in (Sánchez, et al., 2012b), simplifies the matching problem but prevents from applying similarity measures other than edge-counting ones, since many potentially shared taxonomic subsumers will not be discovered.

In order to overcome the above problems, our approach aims to evaluate and discover *all* the semantically equivalent or very similar subsumers of the compared concepts across different ontologies in order to enable an accurate estimation of feature-based similarity. For this aim, we do not solely rely on terminological matching, but also on semantic and structural evidences of similarity that are evaluated coherently with the taxonomic structure. In the following paragraphs, we discuss some aspects of ontology modelling/integration that have been considered during the design of our method:

1. The first aspect defines the scope of the matching problem. Since *semantic similarity* measures the degree of taxonomic resemblance between concepts and, for feature-based similarity measures, this is based on the evaluation of common and non-common subsumers, only these subsumers are needed to be matched in a multi-ontology setting. Due to concept subsumers usually account for a few dozens, rather than hundreds or thousands of concepts that are modelled in large ontologies (Fellbaum, 1998; Nelson, et

al., 2001), the complexity of the matching problem is simplified, both from the scalability point-of-view, and regarding semantic inconsistencies that may appear between ontologies designed by different knowledge engineers (Rodríguez & Egenhofer, 2003).

2. The second aspect is related to the structure of the ontology with respect to the knowledge engineer's point of view. In order to integrate different ontological knowledge, one has to consider that ontologies may represent the same knowledge in significantly different ways (Euzenat & Shvaiko, 2007). This comes from the fact that the knowledge representation process is biased by the engineer's point of view, the application in which the ontology will be used and the distributed nature of the ontology development process. In fact, the level of detail, branching factor and the granularity of the inner taxonomic structure, are aspects decided by the knowledge engineer. As a result of this heterogeneity and lack of consensus, ontological concepts rarely match perfectly (e.g. *renal failure* = *kidney failure*). Thus, due to differences in the level of taxonomic detail, a concept in one ontology may correspond to several specialisations on the other one. To illustrate this aspect, Figure 4 shows the taxonomic structure of two concepts belonging to two ontologies for which the similarity should be assessed: *Systemic lupus erythematosus* in MeSH and *Type I diabetes* in WordNet. By analysing the taxonomies, we observe notorious differences in the granularities of taxonomic trees, in addition to the fact that both concepts present multiple inheritance. The level of abstraction is also different in both ontologies. WordNet introduces a larger amount of abstract concepts at the top of the taxonomic tree because of its general scope (i.e. *entity -> abstract entity -> abstraction -> attribute*), whereas MeSH directly focuses on medical terms because it is framed in the biomedical domain. Given that the compared concepts are both *autoimmune diseases* (i.e. one should expect that a number of their generalisations would be equivalent), in this case, a many-to-many matching would be appropriate to adapt the different granularity degrees. Figure 4 shows several hypothetic matchings. For instance one could match *immune system diseases* and *autoimmune diseases* in MeSH with *autoimmune disorder* in WordNet. Also note that, since the compared concepts are similar but not identical, not all of their subsumers should be matched (e.g. *genetic disease*, *polygenetic disease*, *diabetes* and *diabetes mellitus* should not be matched with any of the subsumers of *Systemic lupus*

erythematosus). Hence, the subsumer matching process should support many-to-many (M:N) matchings so that taxonomic structures with different granularity degrees are coherently integrated. Note that the consideration of M:N matchings prevents from applying classical graph matching methods, which usually focus on one-to-one matchings (Conte, et al., 2004).

3. In turn, because knowledge engineers create taxonomic structures which recursively specialise the meaning of concepts, several authors argue that the amount of shared hyponyms between two concepts give indications of their degree of similarity (Pirró & Seco, 2008; Sánchez, et al., 2012b). As proposed in (Sánchez, et al., 2012b), this argument can be extended to also consider the neighbourhood of the compared subsumers. Therefore, if direct hypernyms and hyponyms of the compared subsumers also share an amount of hyponyms (i.e. they are similar), the evidences of subsumers to be equivalent increase. The exploitation of these implicit evidences of similarity (both semantic and structural) helps to overcome the limitations of approaches based solely on terminological matching (Sánchez, et al., 2012b).
4. Regarding the subsumer matching process, as suggested in (Batet, et al., 2013), the subsumption relation defined between a concept and its hyponyms suggests that, if we have evidences that two hyponyms are equivalent, the evidences on the equivalence of their subsumers should increase. This relation represents the cornerstone of our matching process. Hence, the matching between subsumers can be defined as an iterative process in which matchings discovered in previous stages positively influence further matchings. Specifically, the matching algorithm that we propose iteratively updates matching costs between subsumers by considering matchings and costs computed in earlier steps.

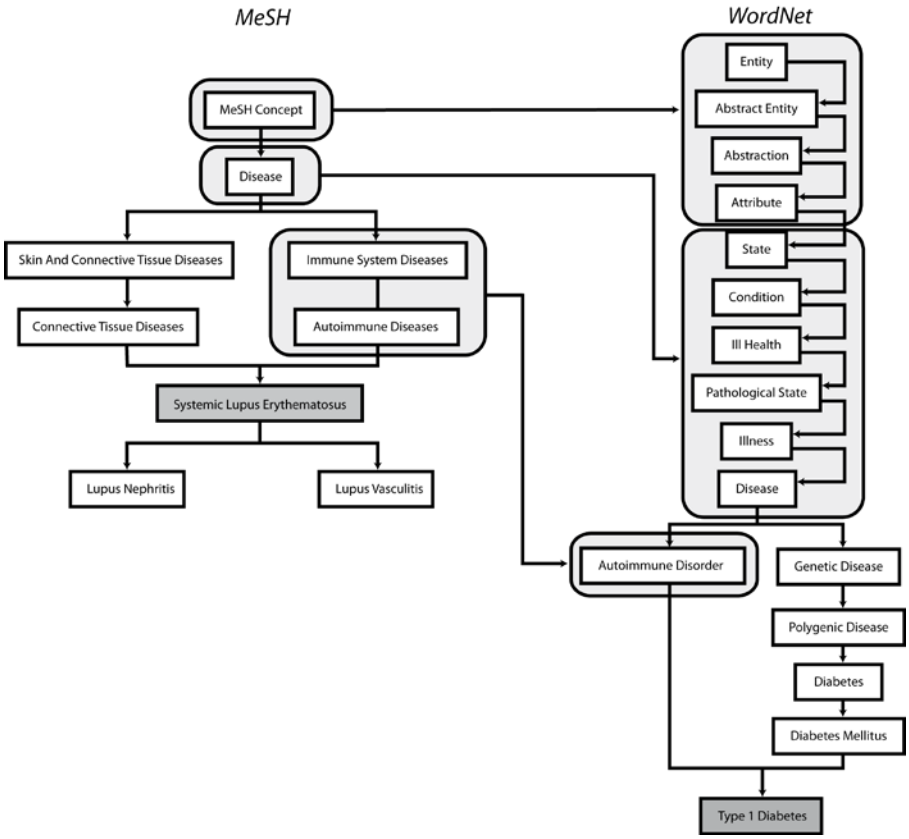

Figure 4. Knowledge structure for *Lupus erythematosus* (in MeSH) and *Type I diabetes* (in WordNet) showing a semantically coherent matching between their subsumers.

4. Our proposal

The aim of our subsumer matching strategy is to improve similarity assessments between concepts belonging to different ontologies focusing on feature-based similarity measures. In this section, we present a method to provide an accurate and semantically coherent matching between concept subsumers. The algorithm relies upon the aspects discussed in section 3. In a nutshell, by considering input ontologies as directed graphs, we define a function to quantify the cost of matching subsumer pairs and propose a subsumer-matching algorithm that seeks for the best set of possible matchings by iteratively updating these costs according to previously discovered matchings.

The following subsections are organised as follows. Subsection 4.1 introduces several definitions used along the section. Subsection 4.2 describes the matching framework together with the cost function used to evaluate the cost of matching two subsumers. Finally, subsection 4.3 presents the algorithm that computes the set of equivalent subsumers.

4.1. Basic definitions

Definition 1. Let C be the set of concepts of an ontology O. We define *concept subsumption* ($<^n$) as a parameterised binary relation $<^n : C \times C$, where n is the number of taxonomic links between a concept and its subsumers. Hence, having two concepts c_i and c_j, $c_i <^n c_j$ implies that c_j is the n^{th} taxonomic subsumer of c_i and, inversely, c_i is the n^{th} specialisation of c_j. □

Particularly, $c_i <^1 c_j$ means that c_i is a *direct hyponym* of c_j; and hence, that c_j is a *direct subsumer* of c_i. Note that $c_i <^0 c_j$ is fulfilled if $c_i = c_j$.

Definition 2. Let $T^{-n}(c) = \{s \in C \mid c <^n s\}$ be the set of *subsumers* of c, which are separated n taxonomic links from it. □

Definition 3. Let $T^{+n}(c) = \{h \in C \mid h <^n c\}$ be the set of *hyponyms* of c, which are separated n taxonomic links from it. □

As a special case, note that $T^0(c) = \{c\}$.

Definition 4. Let $T^{-1}(c) = \{s \in C \mid c <^1 s\}$ be the set of *direct subsumers* of a concept c. □

Definition 5. Let $T^{+1}(c) = \{h \in C \mid h <^1 c\}$ be the set of *direct hyponyms* of a concept c. □

To illustrate the above definitions, consider the MeSH taxonomy of Figure 4. See that, *systemic lupus erythematosus* $<^3$ *disease* , T^{-2}*(systemic lupus erythematosus)={skin and connective tissue disease, immune system disease}*and T^{+1}*(systemic lupus erythematosus)={Lupus Nephritis, Lupus Vasculitis}.*

Definition 6. The *closure* of the subsumption relation ($<^*$) is the union of the results of applying $<^n$ for n=0… d, being d the maximum number taxonomic links between a concept and the *root* node. Consequently, $c_i <^* c_j$ is fulfilled if c_j is a subsumer of c_i at *any* level in the taxonomic tree or if $c_i = c_j$. On the contrary, given c_j, c_i is a taxonomic specialisation of c_j at any taxonomic depth. □

Definition 7. The complete set of *subsumers* of a concept c in the ontology O at any taxonomic level is: $T^{-*}(c) = \{s \in C \mid c <^* s\}$. □

Note that if multiple inheritance relations are modelled into the ontology, the whole set of subsumers (through the different taxonomic trees) is considered.

Definition 8. The complete set of *hyponyms* of a concept c in the ontology O at any taxonomic depth is: $T^{+*}(c) = \{h \in C \mid h <^* c\}$. □

Note that when using the closure, the concept *c* is included both in the subsumer and hyponym sets.

To illustrate the closure of the subsumation relation, consider again the MeSH taxonomy of Figure 4. The closure of the subsumption relation (<*) for *systemic lupus erythematosus* is, T^{-*}*(systemic lupus erythematosus)= { systemic lupus erythematosus, connective tissue diseases, autoimmune diseases, skin and connective tissue diseases, immune system diseases, diseases, and MeSH concept }*, which covers all the taxonomic ancestors of *systemic lupus erythematosus* including itself. Likewise, T^{+*} *(systemic lupus erythematosus)={ Lupus Nephritis, Lupus Vasculitis, systemic lupus erythematosus }*.

4.2. Subsumer matching framework

In this section, we present the subsumer matching framework focused in discover all the *suitable equivalent concepts* between the taxonomic subsumers $T^{-*}(c_1)$ and $T^{-*}(c_2)$ of a pair of concepts c_1 and c_2 that belong to two different ontologies O^1 and O^2 ($c_1 \in O^1$ and $c_2 \in O^2$).

To evaluate the degree of equivalence between subsumer pairs of different ontologies, we rely on static and dynamic evidences of resemblance. Static evidences are those that can be gathered from directly comparing input ontological trees. Dynamic evidences correspond to the recursive propagation of hyponym matchings towards their taxonomic ancestors.

4.2.1 The matching matrix

In order to represent the *many-to-many* equivalence relations between subsumers of the given pair of concepts c_1 and c_2, we rely on a matching matrix *M* (see Definition 9 below). Rows index subsumers of c_1 and columns index subsumers of c_2. *One* values of the matrix indicate a suitable equivalence and *zero* values indicate non-suitable one. This is a common approach in the graph matching community (Conte, et al., 2004). Recall that since the scope of the problem only focuses on the subsumers of the given concepts, the matching matrix just considers these elements instead of considering the complete set of concepts in both ontologies.

Notice that in order to obtain coherent feature-based similarity assessments, the determination of equivalences between similar subsumers is as important as detecting the semantically different

subsumers. Hence, our matching solution should admit, on the one hand, *many-to-many* matchings (as evidences of similarity) and, on the other hand, that some subsumers remain *un-matched* (as evidences of dissimilarity). *Many-to-many* matchings model the situation where one subsumer is represented by several subsumers in the other ontology. *Un-matched* subsumers model the situation where a particular subsumer has no equivalence in the other ontology. To represent this second situation, we follow a usual graph-matching method (Justice & Hero, 2006; Riesen & Bunke, 2009; Wong & You, 1985). The approach is based on modelling un-matched nodes through matching them to artificially added ones, usually called *null-nodes*. In this way, it is ensured that all nodes have, at least, one corresponding node on the other graph. This approach clearly eases the formalisation of the error tolerant graph matching problem (Bunke, 1998; Gao, et al., 2010; Neuhaus & Bunke, 2007), since one is able to include, in the approach, a threshold to indicate the minimum amount of evidence required to consider that two elements are equivalent. We will elaborate on that topic in the next section.

Thus, to adapt the approach to the subsumer matching problem, we define *null-subsumers* as subsumers that they neither contain information, nor they have a taxonomic relation with original subsumers of the ontology. We identify these subsumers with the symbol $\varnothing$. To consider them in the matching process, we extend $T^{-*}(c_1)$ and $T^{-*}(c_2)$ with *null-subsumers* to be of order $\left|T^{-*}(c_1)\right|+\left|T^{-*}(c_2)\right|$ (i.e. a total of $|T^{-*}(c_2)|$ *null-subsumers* are added to $T^{-*}(c_1)$ and vice-versa). We denote these new sets with $\hat{T}^{-*}(c_1)$ and $\hat{T}^{-*}(c_2)$. Accordingly, we formally define the matching matrix as:

Definition 9. Given a pair of elements $s_a \in \hat{T}^{-*}(c_1)$ and $s_i \in \hat{T}^{-*}(c_2)$, we define the matching matrix $M \in \{0,1\}^{\left|\hat{T}^{-*}(c_1)\right|\times\left|\hat{T}^{-*}(c_2)\right|}$, which indicates suitable equivalence between subsumers, as:

$$M[s_a, s_i] = \begin{cases} 1 & if\ s_a \equiv s_i \\ 0 & otherwhise \end{cases} \quad (1)$$

where the relational operator "$\equiv$" indicates whether elements s_a and s_i are conceptually similar (according to the below-described method).

□

As discussed above, and coherently with the usual approach (Gautama, et al., 2006; Justice & Hero, 2006; Riesen & Bunke, 2009), in order to enforce all elements of $\hat{T}^{-*}(c_1)$ and $\hat{T}^{-*}(c_2)$ to be

matched to at least one element of the other set, solutions of the problem must fulfil the following restrictions:

$$\sum_{s_j \in \hat{T}^{-*}(c_2)} M\left[s_a, s_j\right] \geq 1$$
$$\sum_{s_b \in \hat{T}^{-*}(c_1)} M\left[s_b, s_i\right] \geq 1 \quad (2)$$

Note that we model *many-to-many* equivalences and, thus, our restrictions differ slightly of the usual one-to-one approach (Gautama, et al., 2006; Justice & Hero, 2006; Riesen & Bunke, 2009) and contain *greater than* relations instead of *equality*.

4.2.2 Cost between two subsumers

In order to assess the degree of equivalence between two subsumers, we define a cost measure between them. Since *null-subsumers* do not contain neither semantic nor structural information, we define the cost function $f_s : \hat{T}^{-*}(c_1) \times \hat{T}^{-*}(c_2) \rightarrow \{\mathbb{R}, \infty\}$, that differentiates between the three types of equivalence relations: *subsumer-to-subsumer* matching, *null-subsumer-to-subsumer* matchings and *null-subsumer-to-null-subsumer* matching. Formally:

Definition 10. Given a pair of elements $s_a \in \hat{T}^{-*}(c_1)$ and $s_i \in \hat{T}^{-*}(c_2)$ .We define the cost function to evaluate their dissimilarity as:

$$f_s\left(s_a, s_i\right) = \begin{cases} d(s_a, s_i) & if\ s_a, s_i \neq \varnothing \\ K & if\ \left(s_a \neq \varnothing \wedge s_i = \varnothing\right) \vee \left(s_a = \varnothing \wedge s_i \neq \varnothing\right) \\ \infty & if\ s_a \wedge s_i \neq \varnothing \end{cases} \quad (3)$$

□

$d(s_a, s_i)$ represents the *subsumer-to-subsumer* cost, which considers semantic and structural information. Function *d* will be defined in the next section. *K* represents a threshold below which an equivalence relation between two subsumers is considered suitable. Thus, if $d(s_a, s_i)$ for a particular s_a is greater than *K* for all other elements $s_i \in T^{-*}(c_2)$ then, it is considered that subsumer s_a does not have any equivalent subsumer in $T^{-*}(c_2)$ and, thus, it should be matched with a *null-concept* of $\hat{T}^{-*}(c_2)$. On the contrary, if $\exists\, s_i \,|\, d(s_a, s_i) < K$ then the two subsumers are considered equivalent and should be matched. Eventually, since we are not interested in matchings

between *null-concepts*, we forbid them by imposing an infinite cost. From this function, we can define the cost matrix as follows:

Definition 11. The cost matrix $C \in \{\mathbb{R}, \infty\}^{|\hat{T}^{-*}(c_1)| \times |\hat{T}^{-*}(c_2)|}$, which contains the cost of matching subsumers is defined as:

$$C[s_a, s_i] = f_s(s_a, s_i) \tag{4}$$

Given that the degree of overlapping between different ontologies may vary, and the cost function *d* will be based on evaluating that overlap (as detailed in the next section), values of *d* may result in different ranges of costs depending on the given ontologies. Moreover, since the similarity between concepts is not linear to the amount of their evidences of equivalence gathered from the ontologies (Lemaire & Denhière, 2006), neither is our cost function *d*.

As we saw above, threshold *K* and cost function *d* are closely related, since *K* value should be defined coherently with the range of values provided by *d*. Consequently, we model parameter *K* as a function of *d* and a parameter *p* that configures the necessary degree of non-linearity to evaluate similarity evidences:

$$K = Tsd(d, p) := \frac{1 - e^{-p \cdot d(s_a^*, s_i^*)}}{1 - e^{-p}} \tag{5}$$

where,

$$(s_a^*, s_i^*) = \underset{s_a \in T^{-*}(c_1),\, s_i \in T^{-*}(c_2)}{\arg\min} d(s_a, s_i)$$

$$d(\bullet, \bullet) \in [0,1],\ p \in (0, \infty]$$

The value (s_a^*, s_i^*) corresponds to the pair of subsumers of c_1 and c_2 that have minimum distance (according to the *arg min* operator). Parameter *p* takes values in the interval $(0, \infty]$. For values close to zero, function *Tsd* (Threshold), which states the value of constant *K*, resembles a linear function and, for large values, it represents the step function.

In the following two subsections, we will describe, the definition of cost function *C* and cost *d* in order to consider static (subsection 4.2.3) and dynamic (subsection 4.2.4) evidences of resemblance.

4.2.3 Static subsumer matching cost

Given the arguments discussed in section 3 (item 3), we define the cost function d between two subsumers as in (Sánchez, et al., 2012b), according to their degree of *semantic* and *structural* overlapping. The former is related to the local evidences of similarity between two subsumers. The latter, which is addressed to complement these local evidences, is related to the structural resemblances between ontologies.

The *semantic* overlapping cost d_s between two subsumers is computed according to the number of hyponyms they share. It is usual to consider that the set of hyponyms of a concept defines its meaning and helps to differentiate the concept from other ones (Seco, et al., 2004). Thus, a degree of commonality between hyponyms of two concepts gives evidences of their similarity (Sánchez, et al., 2012b). Formally, given two subsumers s_a and s_i, the cost measure is computed as the inverse of the Ochiai coefficient between the amount of shared hyponyms:

$$d_s(s_a, s_i) = 1 - \frac{|T^{+*}(s_a) \cap T^{+*}(s_i)|}{\sqrt{|T^{+*}(s_a)| \cdot |T^{+*}(s_i)|}} \tag{6}$$

Equivalently to (Sánchez, et al., 2012b), we include the *structural* overlapping cost in the cost measure in order to consider the structural resemblance of subsumers' neighbourhood. Formally:

$$d(s_a, s_i) = \frac{1}{2} d_s(s_a, s_i) + \frac{1}{2} \left(\begin{array}{l} \frac{1}{2} \frac{1}{\left| T^{-1}(s_a) \right|} \sum_{\forall s_b \in T^{-1}(s_a)} \min_{\forall s_j \in T^{-1}(s_i)} d_s(s_b, s_j) + \\ + \frac{1}{2} \frac{1}{\left| T^{+1}(s_a) \right|} \sum_{\forall h_b \in T^{+1}(s_a)} \min_{\forall h_j \in T^{+1}(s_i)} d_s(h_b, h_j) \end{array} \right) \tag{7}$$

The cost in (7) is composed by the addition of three terms. The first term is defined as in (6). The second term depends on the local cost of matching subsumers' direct ancestors, $T^{-1}(s_a)$ and $T^{-1}(s_i)$, and the third term depends on the local cost of matching subsumers' direct hyponyms, $T^{+1}(s_a)$ and $T^{+1}(s_i)$. Second and third terms are computed as the mean of the best correspondences between concepts in the adjacent neighbourhood of s_a and s_i (Sánchez, et al., 2012b).

4.2.4 Dynamic subsumer matching cost

Given the taxonomic relation between subsumers, known evidences of equivalence may disclose other semantically equivalent subsumers at higher taxonomic levels (as stated in section 3 (item 4) and suggested in (Batet, et al., 2013)). In order to consider these known evidences, which are assumed to be reliable, we propose a method that, by considering the known equivalences between elements in $\hat{T}^{-*}(c_1)$ and $\hat{T}^{-*}(c_2)$, updates the matching costs between subsumers.

For illustrative purposes consider a simple case. Let C^1 $[s_a,s_i]$ be the cost matrix defined as in (4), which contains only static knowledge that represents the costs of matching subsumers in $s_a \in \hat{T}^{-*}(c_1)$ and $s_i \in \hat{T}^{-*}(c_2)$. In addition, let M^1 be the matching matrix computed with some method (see the following section), and by considering the cost matrix C^1, e.g. taking the matching with least cost. Hence, we could generate a new cost matrix C^2 between elements $s_a \in \hat{T}^{-*}(c_1)$ and $s_i \in \hat{T}^{-*}(c_2)$ by considering its current cost in C^1 and the cost of the matched elements in $h_b \in \hat{T}^{+*}(s_a)$ and $h_j \in \hat{T}^{+*}(s_i)$. We propose to compute this new cost function by averaging the current cost (static component) and the mean cost of its matched hyponyms (dynamic component). Formally,

$$C^2[s_a,s_i] = f_d\left(M^1, C^1, s_a, s_i\right) =$$

$$= \begin{cases} \overbrace{\frac{1}{2}C^1[s_a,s_i]}^{\text{Static component}} + \frac{1}{2}\overbrace{\frac{\sum_{(h_b,h_j)\in B} C^1[h_b,h_j]}{|B|}}^{\text{Dynamic component}} & \text{if } |B|>0 \\ C^1[s_a,s_i] & \text{otherwise} \end{cases} \qquad (8)$$

$$\forall s_a \in \hat{T}^{-*}(c_1), s_i \in \hat{T}^{-*}(c_2)$$

where set B represents the matched hyponyms below nodes s_a and s_i. Formally, the set B is defined as follows:

$$B = \left\{ (h_b,h_j) \middle| h_b \in \left\{ \hat{T}^{+*}(s_a) \setminus s_a \right\} \wedge h_j \in \left\{ \hat{T}^{+*}(s_i) \setminus s_i \right\} \wedge M^1[h_b,h_j] = 1 \right\} \qquad (9)$$

It is worth noting that the cost of matching a subsumer to a null-subsumer is never altered since null-subsumers do not contain taxonomic relationships and, thus, they will never fulfil the relation given in (9). In addition, in order to decrease the cost of a pair of subsumers (i.e. increase

evidences of similarity) in (8), its current cost (static component) is averaged with the mean cost of its matched hyponyms (dynamic component).

Given the cost function in (8), we can define a recursive formulation of the dynamic subsumer matching cost where, at iteration *t+1*, the cost function C^{t+1} depends on the cost function C^t and the matchings M^t computed at iteration *t*.

Definition 12. Given the cost function C^t and the matching matrix M^t, we define the new cost function between elements $s_a \in \hat{T}^{-*}(c_1)$ and $s_i \in \hat{T}^{-*}(c_2)$ at iteration *t*+1 as:

$$C^{t+1}[s_a, s_i] = f_d\left(M^t, C^t, s_a, s_i\right) =$$

$$= \begin{cases} \overbrace{\frac{1}{2} C^t[s_a, s_i]}^{\text{Static component}} + \overbrace{\frac{1}{2} \frac{\sum_{(h_b, h_j) \in B} C^t[h_b, h_j]}{|B|}}^{\text{Dynamic component}} & \text{if } |B| > 0 \\ C^t[s_a, s_i] & \text{otherwise} \end{cases} \quad (10)$$

$$\forall s_a \in \hat{T}^{-*}(c_1), s_i \in \hat{T}^{-*}(c_2)$$

$$B = \left\{ (h_b, h_j) \middle| h_b \in \left\{ \hat{T}^{+*}(s_a) \setminus s_a \right\} \wedge h_j \in \left\{ \hat{T}^{+*}(s_i) \setminus s_i \right\} \wedge M^t[h_b, h_j] = 1 \right\}$$

□

4.3 Subsumer matching algorithm

Given the subsumer matching framework described in subsection 4.2.1 and 4.2.2 and the static and dynamic cost functions proposed in sections 4.2.3 and 4.2.4, and given two concepts $c_1 \in O^1$ and $c_2 \in O^2$, the following iterative algorithm discovers their equivalent subsumers iteratively until a convergence criterion is fulfilled.

Algorithm 1. Algorithm to determine the equivalences between subsumers of two concepts from different ontologies.

1 **Algorithm** compute_subsumer_correspondences(O^1,O^2,c_1,c_2)
2 $M^0[s_a,s_i]:=0, \ \forall \ s_a,s_i$
3 K:=$Tsd(d,p)$
4 t:=1
5 $C^t[s_a,s_i]:= f_s(s_a,s_i,K), \ \forall \ s_a,s_i$ (see (4))
6 *converged*:=*false*
7 **while** (not *converged*) **do**
8 FS:=*extractFeasibleSolutions*(M^{t-1},C^t,K)
9 $M^t[s_a,s_i]:=1$ if $\begin{cases} (s_a, s_i) \in FS \\ \vee \\ M^{t-1}[s_a, s_i] = 1 \end{cases}, \forall\, s_a, s_i$
10 $C^{t+1}[s_a,s_i]:=f_d(M^t,C^t, s_a,s_i), \ \forall \ s_a,s_i$ (see (10))
11 *converged*:=*checkConvergence*(M^t) (see(2))
12 t:=t+1;
13 **end while**
14 **return** M^{t-1}

The proposed method is outlined in Algorithm 1. In the initialisation phase, K value is set by considering, on the one hand, the cost function d and, on the other hand, the desired degree of non-linearity p (as introduced in section 4.2.2). In addition, the cost (line 5) between subsumer pairs is initialised as in (4); in the initial cost only static knowledge is considered because there is still not any evidence of matching. The iterative phase is based on three steps.

1. In the first step (line 8) the function *extractFeasibleSolutions* retrieves, from the non-previously assigned subsumers, the set FS of subsumer pairs that have enough evidences to be considered equivalent (i.e. these are the pairs of subsumers whose cost is lower than value K). In case that no pairs of subsumers have a cost below K, a pair containing a subsumer and *null-subsumer* is taken.
2. In the second step (line 9), these equivalences are reflected in the matching matrix. The new matching matrix, M^t, will contain the assignations of the previous one, M^{t-1}, plus the new matchings found in the current iteration (matchings indicated by set FS).

3. In the third step (line 10), the new cost C^{t+1} of matching subsumers is re-computed by considering the current matching matrix M^t and the cost matrix C^t (see section 4.2.4).

The algorithm repeats steps one, two and three until convergence. This convergence is evaluated by function *checkConvergence*. Specifically, the algorithm continues until all subsumers of $T^{-*}(c_1)$ and $T^{-*}(c_2)$ have been assigned to, at least, one element of the other set $\hat{T}^{-*}(c_2)$ and $\hat{T}^{-*}(c_1)$ (see (2) in section 4.2.1). Note that, if a subsumer in $T^{-*}(c_1)$ is matched with a *null-subsumer* , it will have no equivalent subsumer in $T^{-*}(c_2)$.

5. Evaluation

The suitability of the proposed method has been tested by means of several feature-based measures. Particularly, our ontology matching method has been tested by quantifying the accuracy of these measures when similarity is assessed from the integrated ontological structures provided by our approach. In order to enable an objective evaluation, the accuracy of similarity assessments has been quantified by comparing them with human judgements of similarity for a set of widely-used benchmarks of word pairs. As ontologies, two standard knowledge repositories (WordNet and MeSH) have been used.

5.1 Similarity measures

As stated in the introduction, feature-based measures aim at overcoming some of the limitations of edge-counting methods by evaluating the degree of overlapping and disjunction between ontological features. Thus, in a multi-ontology setting, they can take advantage of the multiple $M:N$ matchings that our method is able to discover.

Since this work focuses on taxonomic *similarity*, we have centred the evaluation on feature-based measures that solely exploit taxonomic knowledge. Formally, given a pair of concepts (c_1, c_2), these measures are based on several well-known similarity coefficients that evaluate similarity according to the degree of overlapping between the set of subsumers of the compared concepts (i.e. $T^{-*}(c_1)$ and $T^{-*}(c_2)$) (Maedche & Zacharias, 2002; Sánchez, et al., 2012a). The following measures have been considered to evaluate our method:

$$Jaccard\_similarity(c_1, c_2) = \frac{|T^{-*}(c_1) \cap T^{-*}(c_2)|}{|T^{-*}(c_1) \cup T^{-*}(c_2)|} \quad (11)$$

The Jaccard coefficient is one of the most commonly applied measures to compute semantic similarity (Maedche & Zacharias, 2002; Sánchez & Batet, 2011). It computes the ratio between common features (i.e. intersection between subsumer sets) normalised by the complete set of features. A similar approach is presented in (Sánchez, et al., 2012a), in which authors evaluate, in a non-linear way, the similarity between concepts as the opposite of the amount of non-shared subsumers:

$$Sanchez_{etal}\_similarity(c_{1,}c_2) = -\log_2\left(1 + \frac{\left|T^{-*}(c_1) \cup T^{-*}(c_2)\right| - \left|T^{-*}(c_1) \cap T^{-*}(c_2)\right|}{\left|T^{-*}(c_1) \cup T^{-*}(c_2)\right|}\right) \quad (12)$$

Other well-known similarity coefficients that have also been used to measure the similarity (Sánchez & Batet, 2011) can be differentiated by the calculus of the normalising factor:

$$Dice\_similarity(c_1, c_2) = \frac{2*|T^{-*}(c_1) \cap T^{-*}(c_2)|}{|T^{-*}(c_1)| + |T^{-*}(c_2)|} \quad (13)$$

$$Ochiai\_similarity(c_1, c_2) = \frac{|T^{-*}(c_1) \cap T^{-*}(c_2)|}{\sqrt{|T^{-*}(c_1)| \cdot |T^{-*}(c_2)|}} \quad (14)$$

5.2 Evaluation criterion, ontologies and datasets

We have reproduced the multi-ontology evaluation scenario proposed in (Al-Mubaid & Nguyen, 2009; Sánchez, et al., 2012b). We tested our method with two significantly different (but with certain degree of overlapping) ontologies: WordNet and MeSH. The former is a lexicon that defines more than 100,000 general concepts of different domains, which are semantically structured in a taxonomic way. The latter consists in a hierarchy of medical and biological terms defined by the U.S. National Library of Medicine. It focuses on indexing clinical documents through more than 22,000 medical concepts. The fact that ontologies with significantly different scopes are considered (i.e. WordNet is a general-purpose repository, whereas MeSH is a medical-specific one), configures a challenging evaluation scenario. Particularly, since taxonomic trees

tend to be quite different among the two ontologies, even for similar concept pairs, the chance of finding terminological matchings between subsumers tends to be low (Batet, et al., 2011).

In such a multi-ontology scenario, the accuracy of the similarity assessments provided by the measures detailed in the above section directly depends on the adequacy of concept matchings discovered among the two ontologies. Hence, by quantifying the accuracy of similarity assessments, we also test the adequacy of our matching method.

In order to evaluate similarity assessments, these are usually compared against similarity judgements provided by human experts. As a result, similarity accuracy is computed as the *correlation* between human ratings and those obtained from computerised methods for a given set of concept pairs. In this work we have used the Pearson's correlation coefficient (14), which has been commonly employed in the literature to evaluate similarity measures (Al-Mubaid & Nguyen, 2009; Batet, et al., 2013; Bollegala, et al., 2009; Pirró, 2009; Sánchez & Batet, 2011; Sánchez & Batet, 2012):

$$Correlation_{Pearson}(sim_{human}, sim_{computerised}) = \frac{\text{cov}(sim_{human}, sim_{computerised})}{\sigma_{sim_{human}} \sigma_{sim_{computerised}}} \quad (15)$$

where sim_{human} stands for similarity ratings provided by human experts for a set of term pairs, $sim_{computerised}$ corresponds to ontology-based similarity assessments for the same set of pairs, $cov(\cdot,\cdot)$ is the covariance between both ratings, and σ is the standard deviation of each set of ratings. A correlation value near 1 indicates that both ratings are very close and, hence, that the computerised assessment accurately reflects human judgements of similarity. On the contrary, correlation values near 0 state that both ratings are almost independent, which indicates a very poor assessment of similarities.

As evaluation benchmarks, several sets of concept pairs, whose similarity were assessed by human experts, exist for the mono-ontology scenario (Hliaoutakis, et al., 2006; Miller & Charles, 1991; Pedersen, et al., 2007; Rubenstein & Goodenough, 1965). In the multi-ontology setting, related works (Al-Mubaid & Nguyen, 2009; Sánchez, et al., 2012b) also use some of these benchmarks by artificially considering that *each* of the two terms of each pair belongs to a *different* ontology.

By following the same methodology, we used the evaluation datasets detailed in (Sánchez, et al., 2012b), which consist of two sets of clinical terms that can be found both in WordNet and MeSH. The first one (named *Dataset1*) covers 20 word pairs taken from (Pedersen, et al., 2007) with averaged similarity ratings provided by different groups of experts from the Mayo Clinic: a group of 3 *physicians*, a group of 9 medical *coders* and the averaged ratings of *both* of them (see Table 1). The second one (named *Dataset2*) consists of 35 medical term pairs taken from (Hliaoutakis, et al., 2006) with averaged similarity ratings provided by 8 medical experts. The multi-ontology scenario has been set as in (Sánchez, et al., 2012b): for each pair of terms, one is considered to be included in WordNet whereas the other one is assumed to be included in MeSH (see the correspondence between terms and ontologies in Table 1 and 2).

Notice that the fact that *all* term pairs in *Dataset1* and *Dataset2* can be found both in WordNet and MeSH also allows computing similarities in a mono-ontology setting (i.e. when term pairs are evaluated solely on WordNet or MeSH). This enables a better interpretation of similarity results obtained in the multi-ontology scenario described above, because its accuracy depends on the ground truth accuracies that can be achieved from *each* of the individual ontologies. For this reason, mono-ontology results approximately *bind* the best expected correlation that a multi-ontology strategy may achieve for the same ontologies.

Table 1. *Dataset1*: each row shows the correspondence between terms in the multi-ontology scenario, the similarity ratings for two types of experts and the average of both of them (Pedersen, et al., 2007).

WordNet term	*MeSH term*	*Physician ratings*	*Coder ratings*	*Both*
Renal failure	Kidney failure	4.0	4.0	4.0
Myocardium	Heart	3.3	3.0	3.15
Infarct	Stroke	3.0	2.8	2.9
Abortion	Miscarriage	3.0	3.3	3.15
Schizophrenia	Delusion	3.0	2.2	2.6
Adenocarcinoma	Metastasis	2.7	1.8	2.25
Stenosis	Calcification	2.7	2.0	2.35
Diarrhoea	Stomach cramps	2.3	1.3	1.8
Atrial fibrillation	Mitral stenosis	2.3	1.3	1.8
Rheumatoid arthritis	Lupus	2.0	1.1	1.55
Osteoarthritis	Carpal tunnel syndrome	2.0	1.1	1.55
Hypertension	Diabetes mellitus	2.0	1.0	1.5
Acne	Syringe	2.0	1.0	1.5
Antibiotic	Allergy	1.7	1.2	1.45
Multiple sclerosis	Psychosis	1.0	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0	1.0
Cellulitis	Depression	1.0	1.0	1.0
Hyperlipidaemia	Metastasis	1.0	1.0	1.0

Table 2. *Dataset2*: each row shows the correspondence between terms in the multi-ontology scenario, and similarity ratings provided by human experts (Hliaoutakis, et al., 2006).

WordNet term	*MeSH term*	*Expert ratings*
Appendicitis	Anemia	0.031
Otitis Media	Infantile Colic	0.156
Dementia	Atopic Dermatitis	0.060
Malaria	Bacterial Pneumonia	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156
Antibacterial Agents	Amino Acid Sequence	0.155
Congenital Heart Defects	Acq. Immunno. Syndrome	0.060
Meningitis	Tricuspid Atresia	0.031
Sinusitis	Mental Retardation	0.031
Hypertension	Kidney Failure	0.500
Hyperlipidemia	Hyperkalemia	0.156
Hypothyroidism	Hyperthyroidism	0.406
Sarcoidosis	Tuberculosis	0.406
Vaccines	Immunity	0.593
Asthma	Pneumonia	0.375
Diabetes Mellitus	Diabetic Nephropathy	0.500
Lactose Intolerance	Irritable Bowel Syndrome	0.468
Urinary Tract Infection	Pyelonephritis	0.656
Sepsis	Neonatal Jaundice	0.187
Anemia	Deficiency Anemia	0.437
Psychology	Cognitive Science	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Infarction	Myocardial Ischemia	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary Stenosis	Aortic Stenosis	0.531
Breast Feeding	Lactation	0.843
Antibiotics	Antibacterial Agents	0.937
Seizures	Convulsions	0.843
Ache	Pain	0.875
Malnutrition	Nutritional Deficiency	0.875
Measles	Rubeola	0.906
Chicken Pox	Varicella	0.968
Down Syndrome	Trisomy 21	0.875

5.3 Subsumer matching strategies

In order to compare the practical benefits that our method provides for feature-based similarity assessment against related works, we have also implemented and evaluated other subsumer matching strategies found in the literature:

- *Root matching (RM)*. Root nodes of the two ontologies are joined (Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003). All other subsumers will be considered as different.
- *Terminological matching (TM)*. Only subsumer pairs with the same textual label are matched (Al-Mubaid & Nguyen, 2009; Saruladha, et al., 2010).
- *Terminological subsumption (TS)*. The most specific pair of terminologically equivalent subsumers is found and it is considered as the least common subsumer (LCS). Then, more general subsumers are considered equivalent (since the subsumption relation is assumed to be propagated upwards), whereas more specific ones remain disjoint, as proposed in (Batet, et al., 2013).
- *Semantic subsumption (SS).* The approach proposed in (Sánchez, et al., 2012b) to detect similar - but not necessarily terminologically identical - subsumers is used to detect the most specific pair of equivalent subsumers (i.e. the LCS). By using the same strategy as in the *terminological subsumption* case, more general subsumers are assumed to be equivalent, whereas more specific ones will be considered as different.

Regarding the implementation of our proposal, the parameter p was empirically chosen to be 0.3 because the difficulty of knowing, a priory, the degree of non-linearity required in the similarity measure. This value of p makes the definition of threshold K (5) almost a linear function (see section 4.2.2).

5.4 Results and discussion

Table 3 details the correlation values obtained for each of the feature-based measures introduced in section 5.1 (Sanchez et al.'s, Jaccard's, Dice's and Ochiai's coefficients) for *i)* a mono-ontology setting in which all pairs are evaluate solely on WordNet or MeSH, and *ii)* a multi-ontology setting in which each term of each pair is evaluated in a different ontology, as detailed in Tables 1 and 2. In this latter case, several subsumer matching strategies have been implemented in addition to *ours*, as detailed in section 5.3 (i.e. *root matching* (RM), *terminological matching*

(TM), *terminological subsumption* (TS) and *semantic subsumption* (SS)). Results are shown for the two benchmarks and ratings introduced in section 5.2 (i.e. *physician's, coder's* and *both* ratings for *Dataset1* and expert ratings for *Dataset2*).

Table 3. Correlation values obtained for the different measures, ontology combinations and matching strategies for the two evaluated benchmarks. Rows in **boldface** show the results of our proposal whereas rows in *italics* show approximate correlation boundaries provided by the mono-ontology scenario.

Similarity Measure	*Ontologies*	*Subsumer matching strategy*	*Dataset 1 Physicians*	*Dataset 1 Coders*	*Dataset 1 Both*	*Dataset2*
Sanchez et al.	*MeSH*	-	*0.64*	*0.74*	*0.72*	*0.76*
Sanchez et al.	*WordNet*	-	*0.53*	*0.63*	*0.60*	*0.64*
Sanchez et al.	MeSH + WordNet	RM	-0.13	-0.01	-0.07	0.02
Sanchez et al.	MeSH + WordNet	TM	0.43	0.31	0.38	0.44
Sanchez et al.	MeSH + WordNet	TS	0.62	0.61	0.64	0.68
Sanchez et al.	MeSH + WordNet	SS	0.57	0.71	0.66	0.71
Sanchez et al.	**MeSH + WordNet**	**Our proposal**	**0.65**	**0.76**	**0.73**	**0.74**
Jaccard	*MeSH*	-	*0.66*	*0.73*	*0.72*	*0.76*
Jaccard	*WordNet*	-	*0.51*	*0.59*	*0.57*	*0.61*
Jaccard	MeSH + WordNet	RM	-0.13	0.01	-0.07	0.02
Jaccard	MeSH + WordNet	TM	0.43	0.31	0.38	0.44
Jaccard	MeSH + WordNet	TS	0.62	0.60	0.63	0.69
Jaccard	MeSH + WordNet	SS	0.53	0.70	0.65	0.70
Jaccard	**MeSH + WordNet**	**Our proposal**	**0.63**	**0.73**	**0.71**	**0.73**
Dice	*MeSH*	-	*0.67*	*0.69*	*0.71*	*0.75*
Dice	*WordNet*	-	*0.44*	*0.50*	*0.49*	*0.55*
Dice	MeSH + WordNet	RM	0.13	-0.01	-0.07	0.02
Dice	MeSH + WordNet	TM	0.42	0.30	0.37	0.47
Dice	MeSH + WordNet	TS	0.61	0.56	0.60	0.68
Dice	MeSH + WordNet	SS	0.54	0.64	0.62	0.68
Dice	**MeSH + WordNet**	**Our proposal**	**0.60**	**0.67**	**0.66**	**0.70**
Ochiai	*MeSH*	-	*0.68*	*0.69*	*0.71*	*0.75*
Ochiai	*WordNet*	-	*0.44*	*0.50*	*0.49*	*0.55*
Ochiai	MeSH + WordNet	RM	-0.12	-0.02	-0.07	0.08
Ochiai	MeSH + WordNet	TM	0.41	0.28	0.35	0.47
Ochiai	MeSH + WordNet	TS	0.60	0.56	0.60	0.68
Ochiai	MeSH + WordNet	SS	0.54	0.65	0.61	0.68
Ochiai	**MeSH + WordNet**	**Our proposal**	**0.60**	**0.66**	**0.66**	**0.70**

The analysis of the correlation values obtained in the mono-ontology scenario shows a significant difference between MeSH and WordNet. The use of the former as the knowledge base

to extract similarity evidences provides more accurate assessments and, hence, higher correlations than WordNet (e.g. from 0.53-0.64 to 0.64-0.76 for Sanchez et al.'s measure). This makes sense because, even though both ontologies contain medical terms, MeSH solely models medical knowledge whereas WordNet has a general focus. Notice that, as stated above, correlation results obtained in these mono-ontology settings approximately *bind* the best expected correlation that a multi-ontology strategy may achieve for the same ontologies (i.e. when one term is evaluated in one ontology and the other term in a different one), which puts the results obtained in the multi-ontology scenario in context. Ideally, one would expect that, in the best case, correlations obtained by means of the subsumer matching method would stay close to the *best* mono-ontology results, which are those of MeSH in these tests.

We observe notable differences between the different subsumer matching strategies implemented by related works. Obviously, the *root matching (RM)* produces the poorest results because, with this strategy, all term pair will appear to be maximally different when computing similarities. Correlation values are very close to zero in all cases, which reflects an almost random assessment. The matching strategy based on *terminological matching (TM)* (i.e. to consider as equivalent *only* those subsumer pairs with identical labels) significantly improves the results even though those are, in all cases, quite below the worst mono-ontology case, that is, those of WordNet (e.g. 0.31-0.44 vs. 0.53-0.64 for Sanchez et al.'s measure). Results obtained with this strategy reflect the amount of terminological equivalences between the two ontologies and establish a baseline to be improved by using more elaborated matching methods. Results obtained with the matching method based on *terminological subsumption (TS)* (i.e. all subsumers above the terminologically equivalent LCS are considered equivalent) and for the *semantic subsumption (SS)* (which relies on an assessment of the most similar subsumer pair to act as LCS) are quite comparable. Both provide correlation values that, in most cases, stay between the two boundaries defined by the mono-ontology results (e.g. from 0.57 to 0.71 for Sanchez et al.'s measure). From the results, we can conclude that it is reasonable to assume that subsumer pairs above the terminologically equivalent LCS should also be considered equivalent, as it is done for the terminological subsumption strategy. The *terminological subsumption* strategy, however, is hampered by the fact that, only in the best case, the detected LCS will be the appropriate one whereas, in all other cases, the terminologically equivalent LCS will be more abstract than desired.

This situation underestimates the degree of similarity between the compared concepts. Even though the *semantic subsumption* method tends to minimise this problem, because it does not solely rely on terminological equivalences, if the selected LCS is excessively specialised, the opposite situation will occur and the similarity between the compared concepts will be overestimated. Both methods share the problem that all subsumers above the LCS are considered equivalent. This may not be the most appropriate decision, especially in cases of multiple inheritance; that is, concepts may share some taxonomic trees of ancestors but not necessarily all of them (see an example in Figure 4).

In any case, correlation values obtained by means of our method reflect a more accurate similarity assessment in comparison with related works (i.e. RM, TM, TS and SS, as detailed in section 5.3) for all the measures and benchmarks. The comparison of our results against those obtained in a mono-ontology scenario shows that our approach closely approximates the *best* results reported in the ideal mono-ontology scenario (i.e. those of MeSH). For example, for *coder's* ratings from *Dataset1* (column 5 from Table 3), we obtain a correlation value of 0.76 for Sanchez et al.'s similarity measure, which improves results of all related works (whose best correlation is 0.71 for the SS matching strategy), and closely approximates to the best mono-ontology result (i.e. 0.74 for MeSH). Given that correlation values range from around 0 (for the root matching strategy) to a maximum of 0.76, these results reflect an approximate improvement of a 7% against related works.

The behaviour of the proposed method is in essence similar to the *terminological* and *semantic subsumption* approaches, and especially similar to the latter because they share the "static" evaluation of ontological similarities (as detailed in section 4.2.3). This is reflected on the fact that a pair of subsumers that have been matched (i.e. the LCS) positively affects the matching cost of their ancestors. As a result, concept pairs above a discovered LCS are likely to be matched. However, in the proposed algorithm, subtle differences between taxonomic ancestors are better captured, thus enabling a more accurate matching process. This accuracy relies on the fact that some subsumer pairs with high matching cost values (e.g. those belonging to semantically different taxonomic trees, in case of multiple inheritance) will remain unmatched. Since this situation gives evidences of dissimilarity, the method proposed in this paper better quantifies the

commonalities and disjunctions between concepts and, hence, it enables a more accurate assessment of similarity.

The analysis of correlation values for each dataset shows that the best results are obtained with ratings provided for *Dataset1* with coder ratings and for *Dataset2*. In the first case, medical coders were specifically trained to provide taxonomically coherent ratings of similarity (Pedersen, et al., 2007). On the contrary, physicians freely rated term pairs. As a result, semantic similarity measures better correlate with coder ratings. For *Dataset2*, it is worth noting the relatively high correlation obtained when relying solely on *terminological matching*. As stated in (Batet, et al., 2013; Sánchez, et al., 2012b), term pairs for this benchmark seem to be selected so that their ancestors terminologically match quite frequently, since the author's method relied on this principle (Hliaoutakis, et al., 2006).

Results for the different feature-based similarity measures are quite consistent and comparable, since all of them are based on the *same* premise: similarity is assessed as the ratio between the number of common and non-common taxonomic subsumers. As a result, all of them evaluate the same amount of similarity/dissimilarity evidences. The only difference is the similarity coefficient used in each case, which tends to produce subtle differences (Sánchez & Batet, 2011). In any case, the approach presented in (Sánchez, et al., 2012a) tends to produce the most accurate assessments.

6. Conclusions

Given the usefulness of semantic similarity, any step towards providing more accurate results or improving its applicability will have a positive effect in many concrete applications. Regarding the former, more precise similarity computation paradigms have been proposed in recent years, such as the feature-based measures, which have reached state-of-the-art results (Sánchez, et al., 2012a). However, these measures are designed to exploit a single ontology. In order to improve their applicability in cases in which the concepts to compare can only be found in *different* ontologies, this paper proposes a general solution to match taxonomic ancestors of different ontologies. On the contrary to simpler methods based solely on terminological matching, our approach considers implicit evidences of equivalence, such as the degree of overlapping between concept hyponyms and those of their neighbourhood (i.e. structural similarity). In addition to these "static" evidences, an iterative algorithm is presented that progressively re-computes subsumer matching costs

according to previously discovered matchings, which is coherent with the notion of taxonomic subsumption. As a result, similarity assessment in a multi-ontology scenario can be improved in comparison with related works. The theoretical advantages of our method have been empirically evaluated by means of several feature-based measures applied to two standard benchmarks and two widely used ontologies. Similarity accuracies obtained when using the proposed method surpassed those resulting from other matching strategies proposed by related works and stay close to the results obtained in an ideal mono-ontology scenario.

As future work, we plan to expand the evaluation of our method to other domains and ontologies. Moreover, evaluations carried on specific tasks like the classification of heterogeneous datasets (Batet, et al., 2010) will illustrate the potential benefits that our method may bring in a practical setting.

Acknowledgements

This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, 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).

References

Al-Mubaid, H., & Nguyen, H.A. 2009. Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 39*, 389-398.

Batet, M. 2011. Ontology based semantic clustering. *AI Communications, 24*, 291-292.

Batet, M., Sánchez, D., & Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, *44*, 118-125.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. 2013. Semantic similarity estimation from multiple ontologies. *Applied Intelligence*, *38*, 29-44.

Batet, M., Valls, A., Gibert, K., & Sánchez, D., 2010. Semantic clustering using multiple ontologies, 13th International Conference on the Catalan Association for Artificial Intelligence. Publishing, pp. 207-216.

Bergamaschi, B., Castano, S., Vermercati, S.D.C.d., Montanari, S., & Vicini, M., 1998. An Intelligent Approach to Information Integration, in: N. Guarino (Ed.), Proceedings of the First International Conference Formal Ontology in Information Systems. Publishing, pp. 253-268.

Blank, A., 2003. Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology, in: R. Eckardt, K. von Heusinger & C. Schwarze (Eds.), Words and Concepts in Time: towards Diachronic Cognitive Onomasiology. Publishing, Berlin, Germany, pp. 37-66.

Bollegala, D., Matsuo, Y., & Ishizuka, M., 2009. A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web, in: P. Koehn & R. Mihalcea (Eds.), Conference on Empirical Methods in Natural Language Processing, EMNLP 2009. Publishing, Singapore, Republic of Singapore, pp. 803–812.

Bunke, H. 1998. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters, 18*, 689-694.

Conte, D., Foggia, P., Sansone, C., & Vento, M. 2004. Thirty Years Of Graph Matching In Pattern Recognition. *International Journal of Pattern Recognition and Artificial Intelligence, 18*, 265-299.

Euzenat, J., & Shvaiko, P. 2007. *Ontology Matching*. Springer Verlag, Amsterdam.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Gangemi, A., Pisanelli, D., & Steve, G., 1998. Ontology Integration: Experiences with Medical Terminologies, in: N. Guarino (Ed.), Formal Ontology in Information Systems. Publishing, pp. 163-178.

Gao, X., Xiao, B., Tao, D., & Li, X. 2010. A survey of graph edit distance. *Pattern Analysis and applications, 13*, 113-129.

Gautama, S., Bellens, R., Tré, G.D., & D'Haeyer, J., 2006. Relevance Criteria for Data Mining Using Error-Tolerant Graph Matching Combinatorial Image Analysis. Publishing, pp. 277-290.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., & Milios, E.E. 2006. Information Retrieval by Semantic Similarity. *International Journal on Semantic Web and Information Systems, 2*, 55-73.

Justice, D., & Hero, A. 2006. A binary linear programming formulation of the graph edit distance. *Transactions on Pattern Analysis and Machine Intelligence, 28*, 1200-1214.

Lemaire, B., & Denhière, G. 2006. Effects of High-Order Co-occurrences on Word Semantic Similarities. *Current Psychology Letters - Behaviour, Brain and Cognition, 18*, 1.

Maedche, A., & Zacharias, V., 2002. Clustering Ontology-Based Metadata in the Semantic Web, 6th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD 2002. Publishing, London, UK, pp. 348-360.

Martínez, S., Sánchez, D., Valls, A., & Batet, M. 2012. Privacy protection of textual attributes through a semantic-based masking method. *Information Fusion, 13*, 304-314.

Mena, E., Kashyap, V., & Sheth, A., 1996. OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies, International Conference of Cooperative Information Systems, CoopIS 1996. Publishing.

Miller, G.A., & Charles, W.G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*, 1-28.

Nelson, S.J., Johnston, D., & Humphreys, B.L., 2001. Relationships in Medical Subject Headings, Relationships in the Organization of Knowledge. Publishing, pp. 171-184.

Neuhaus, M., & Bunke, H., 2007. A Quadratic Programming Approach to the Graph Edit Distance Problem in: G.-B.R.i.P. Recognition (Ed.). Publishing.

Pedersen, T., Pakhomov, S., Patwardhan, S., & Chute, C. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics, 40*, 288-299.

Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. 2006. X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management, 4*, 233-237.

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

Pirró, G., & Seco, N., 2008. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content, in: R. Meersman & Z. Tari (Eds.), OTM 2008 Confederated International Conferences CoopIS, DOA, GADA, IS, and ODBASE 2008. Publishing, Monterrey, Mexico, pp. 1271-1288.

Rada, R., Mili, H., Bichnell, E., & Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 9*, 17-30.

Riesen, K., & Bunke, H. 2009. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision Computing, 27*, 950-959.

Rodríguez, M.A., & Egenhofer, M.J. 2003. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering, 15*, 442–456.

Rubenstein, H., & Goodenough, J. 1965. Contextual correlates of synonymy. *Communications of the ACM, 8*, 627-633.

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

Sánchez, D., & Batet, M. 2012. A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge. *International Journal on Semantic Web and Information Systems, 8*, 34-50.

Sánchez, D., Batet, M., Isern, D., & Valls, A. 2012a. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications, 39*, 7718-7728.

Sánchez, D., Castellà, J., & Viejo, A. 2013. Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines. *Information Sciences, 218*, 17-30.

Sánchez, D., Isern, D., & Millán, M. 2011. Content Annotation for the Semantic Web: an Automatic Web-based Approach. *Knowledge and Information Systems, 27*, 393-418.

Sánchez, D., Solé-Ribalta, A., Batet, M., & Serratosa, F. 2012b. Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics, 45*, 141-155

Saruladha, K., Aghila, G., & Bhuvaneswary, A. 2010. Computation of Semantic Similarity among Cross Ontological Concepts for Biomedical Domain. *Journal of Computing, 2*, 111-118.

Seco, N., Veale, T., & Hayes, J., 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet, in: R. López de Mántaras & L. Saitta (Eds.), 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004. Publishing, Valencia, Spain, pp. 1089-1090.

Weinstein, P., & Birmingham, W.P., 1999. Comparing Concepts in Differentiated Ontologies, 12th Workshop on Knowledge Acquisition, Modeling and Management, KAW 1999. Publishing, Banff, Alberta, Canada.

Wong, A., & You, M. 1985. Entropy and Distance of Random Graphs with Application to Structural Pattern Recognition. *Transaction on Pattern Analysis and Machine Intelligence, PAMI-7*, 599-609.

Wu, Z., & Palmer, M., 1994. Verb semantics and lexical selection, 32nd annual Meeting of the Association for Computational Linguistics. Publishing, Las Cruces, New Mexico, pp. 133 - 138.