Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain

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

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili

Av. Països Catalans, 26, 43007, Tarragona, Catalonia (Spain)

Keywords: semantic similarity, multiple ontologies, MeSH, WordNet

[*] Corresponding author.

E-mail: david.sanchez@urv.cat. Postal address: Av. Països Catalans, 26. 43007 Tarragona, Catalonia (Spain). Telephone: +34 977256563, Fax: +34 977559710.

Abstract

The estimation of the semantic similarity between terms provides a valuable tool to enable the understanding of textual resources. Many semantic similarity computation paradigms have been proposed both as general-purpose solutions or framed in concrete fields such as biomedicine. In particular, ontology-based approaches have been very successful due to their efficiency, scalability, lack of constraints and thanks to the availability of large and consensus ontologies (like WordNet or those in the UMLS). These measures, however, are hampered by the fact that only one ontology is exploited and, hence, their recall depends on the ontological detail and coverage. In recent years, some authors have extended some of the existing methodologies to support multiple ontologies. The problem of integrating heterogeneous knowledge sources is tackled by means of simple terminological matchings between ontological concepts. In this paper, we aim to improve these methods by analysing the similarity between the modelled taxonomical knowledge and the structure of different ontologies As a result, we are able to better discover the commonalities between different ontologies and hence, improve the accuracy of the similarity estimation. Two methods are proposed to tackle this task. They have been evaluated and compared with related works by means of several widely-used benchmarks of biomedical terms using two standard ontologies (WordNet and MeSH). Results show that our methods correlate better, compared to related works, with the similarity assessments provided by experts in biomedicine.

1 Introduction

With the success of the Information Society, the amount of electronic information available has increased greatly in recent years. Biomedicine is a field in which information is of the utmost importance, and in which large electronic databases are needed both for daily and research tasks. By means of intelligent data analysis, large sources of information (such as patient records, visit outcomes, statistical data obtained from daily tasks, etc.) can be analysed to extract new useful knowledge. Much of this information, however, consists of textual resources that are hard to manage due to the lack of textual understanding capabilities of computerised systems.

Computational linguistics provides techniques to enable the understanding of text. The estimation of the *semantic similarity* is one of the most basic tasks. It aims to quantify the similarity between a pair of terms according to their conceptual (i.e. semantic) resemblance, rather than their lexical similarity [1]. In this manner, the meaning of terms is taken into consideration during data analysis, mimicking the way in which humans interpret textual resources. Because semantics is an inherently human feature, semantic similarity approaches rely on predefined knowledge sources containing implicit or explicit evidences on which the similarity assessment can be based. In all cases, similarity is estimated according to the degree of commonality observed in the background sources for the compared terms.

According to the theoretical principles and the type of knowledge source, different families of measures can be identified. On one hand, corpora-based measures rely on unstructured or semi-structured textual corpus (e.g. thesaurus, tagged documents, the Web, etc.) to estimate the similarity between terms as a function of their distributional characteristics [2]. Term co-occurrences are typically taken as the evidence of commonality on which the similarity assessment can be based (e.g. *metastasis* and *cancer* are similar because both terms tend to cooccur) [3]. On the other hand, ontology-based measures compute similarity by mapping input terms to ontological concepts and estimating the similarity by analysing the modelled semantic relationships (usually hyponymy/hypernymy) [4]. Ontologies provide a formal and machine-readable conceptualisation of a domain, by means of a unified terminology and semantic inter-relations [5] and have an outmost importance in the biomedical field [6]. In recent years, many ontologies have been developed, ranging from general-purpose ones (such as WordNet [7]) to domain-specific sources (such as those in the UMLS repository). In ontology-based measures, taxonomical knowledge is typically exploited. Most of them rely on the *common ancestors* of the compared terms as the evidence on which the similarity is based (e.g., *bronchitis* and *flu* are similar because both are specialisations of *disorders of the respiratory system*).

Compared to corpora-based similarity measures, ontology-based ones are characterised by their efficiency (because they only explore semantic networks rather than analysing large corpora) and lack of dependencies on external resources (i.e. representative domain corpora) [8]. Moreover, they are less affected by ambiguity (because unstructured domain corpora contain words rather than concepts) and data sparseness (i.e. the fact that the amount of available corpora is not enough to extract robust conclusions) [3]. Ontology-based measures, however, completely depend on the coverage and detail of the background ontology. If an input term is not found in the ontology, its similarity to another term cannot be assessed. This situation is common when dealing with cross-domain data and also in domains (such as biomedicine) in which concepts are spread through several heterogeneous knowledge sources (e.g. those in the UMLS such as SNOMED CT or MeSH which are created with different scopes and purposes) [9, 10]. This limitation, as acknowledged by several authors [1, 11-13], can be overcome by exploiting *multiple ontologies*.

As it will be shown in section 3, ontology-based related works very rarely support more than one ontology. The main difficulty arises when evaluating a term found in one ontology with another one covered by a different ontology. As stated above, similarity estimation relies on the commonalities between terms. In a multi-ontology scenario, this implies the discovery of common taxonomical ancestors for the compared terms among several ontologies. Related works rely on terminological matchings between taxonomical ancestors of different ontologies (i.e. ancestors with identical labels are matched) [1, 10]. These approaches are hampered by the fact that ontologies rarely model concepts in the same way or refer to them using the same label (due to synonymy). Hence, in many situations, it is not possible to discover equivalent ancestors or the selected ones may not be the most adequate.

In this paper, we propose two methods that overcome the limitations of a strict terminological matching of taxonomical ancestors. The first one, relying on the principles of knowledge representation, considers explicit knowledge modelled in the ontology to estimate the semantic overlapping between taxonomical ancestors of different ontologies. The second one exploits, additionally, the net of semantic links and the structural similarities between several ontologies as an indication of implicit semantics. Both methods aim to quantify the chance that terminologically-different ancestors are equivalent. And thus, they allow discovering more adequate common ancestors for the compared terms among different ontologies and enabling more accurate similarity estimations. Both methods have been evaluated by means of several widely-used benchmarks (consisting of biomedical term pairs) adapted to the multi-ontology scenario. The similarity estimation accuracy obtained for several ontology-based measures shows a noticeable improvement when using our methods instead of those of related works.

The rest of the paper is organised as follows. Section 2 introduces ontology-based methods for semantic similarity assessment. Section 3 reviews strategies proposed by related works to enable the similarity assessment across different ontologies, while section 4 discusses their main limitations. Section 5 presents our approach. It describes how ontologies are semantically and structurally analysed and formalises two methods to enable the similarity assessment across different ontologies. Section 6 presents the evaluation scenario and the results obtained for several benchmarks of biomedical terms. Section 7 discusses the results in comparison with related works. The final section contains the conclusions and some lines of future research.

2 Ontology-based semantic similarity

Ontology-based measures analyse the knowledge modelled in an ontology to assess the similarity between terms (which are mapped to ontological concepts by matching their labels). Several approaches can be distinguished.

Information Content-based approaches assess the similarity between concepts as a function of the Information Content (IC) that both concepts have in common. The common information between two concepts is represented by the IC of their *Least Common Subsumer* (*LCS*) (i.e. the most specific taxonomical ancestor that the two concepts have in common in the ontology) [14-16]. The IC of a concept can be either computed from its probability of occurrence in a corpus (i.e. the more frequent a concept appears in a corpus, the lower its IC will be) [14], or from its degree of taxonomical specialisation in the background ontology (i.e. the larger the number of hyponyms of a concept, the more general its meaning will be and the lower its IC will be) [12, 17, 18]. As stated in the introduction, pure ontology-based approaches, like the latter one, are preferred to corpora-based ones due to their higher scalability.

Feature-based measures estimate the similarity of concepts by analysing the amount of common and non-common knowledge features found in the ontology. Taxonomic and non-taxonomic relationships as well as concept descriptions (i.e. glosses) retrieved from dictionaries are features commonly considered in the assessment of similarity [11, 13, 19]. In these approaches, term commonality (i.e. similarity) is computed as a function of the amount of terminological overlapping between concept features (i.e. related concepts and/or glosses). The main drawback of these measures is that they rely on features such as non-taxonomical relationships or term descriptions which are rarely found in ontologies [20].

Finally, *edge-counting* measures rely on the structural model defined by the taxonomical relationships modelled in the ontology. These measures base the similarity assessment on the length of the *shortest path* separating two concepts, defined by going through taxonomical generalisations modelled in an ontology [4]. Note that the shortest taxonomical path between two concepts is the one that goes through their LCS that, again, represents their commonality. Thanks to their simplicity, edge-counting measures have been widely used in the past in many contexts [8].

In [4], Rada proposed a simple edge-counting measure which quantifies the semantic distance (i.e. the inverse to similarity) of two concepts c_1 and c_2 as the sum of the number of links of c_1 and c_2 to their LCS (*LCS*(c_1,c_2)) (i.e. their minimum taxonomical path) (1).

$$dis_{Rad}(c_1,c_2) = N_1 + N_2 \qquad (1)$$

where N_1 and N_2 are the minimum number of taxonomical links from c_1 and c_2 to their LCS, respectively. Note, that if c_1 is an specialisation of c_2 then LCS(c_1, c_2) = c_2.

Because the above measure produces absolute values that are difficult to compare when they are computed from different ontologies, Leacock and Chodorow [21] (*L&C*) normalised the value by the maximum depth *D* of the taxonomy, evaluating the path length in a non-linear fashion (2).

$$sim_{L\&C}(c_1,c_2) = -\log\left(\frac{N_1 + N_2 + 1}{2 \times D}\right) \qquad (2)$$

In addition to the path length, Wu and Palmer [22] (*W&P*) also took into consideration that the similarity between a concept pair in an upper level of the taxonomy should be lower than the similarity for a pair in a lower level (because the meaning of the latter one is more specialised and, hence, less differentiated). For this reason, they consider the relative depth of the LCS of the concept pairs in the taxonomy as an indication of similarity (3).

$$sim_{W\&P}(c_1,c_2)=\frac{2\times N_3}{N_1+N_2+2\times N_3} \tag{3}$$

where and N_3 is the number of *is-a* relations from the LCS to the root of the ontology. Note that, because the measure is normalised, it ranges from 1 (for identical concepts) to 0.

Note that most ontology-based measures completely rely on evaluating the LCS of the compared concepts as the evidence of commonality on which the similarity assessment is based. Hence, in a multi-ontology scenario (discussed in the next section), in which c_1 and c_2 belong to different ontologies, the discovery of a suitable LCS between ontologies is mandatory to enable similarity estimations.

3 Semantic similarity from multiple ontologies

As noted in [11], when dealing with multiple knowledge sources, the classical approach has been to map/align [23] concepts of different ontologies into a single one [24]. However, the construction of an integrated ontology for similarity estimation is costly if not impractical. Moreover, the integrated ontology may be also hampered by the difficulties of coherently map all ontological concepts due to inconsistencies and mismatches among ontologies. For this reason, more recent related works on semantic similarity estimation from multiple ontologies tackle the problem from a different perspective: they only evaluate the ontological knowledge related to the evaluated concepts to find the commonalities that enable the similarity assessment. On one hand, this minimises the amount of incosistences resulting from a complete integration process, due to only partial and closely related knowledge sctructures are analysed; on the other hand, it provides an scalable solution from the semantic similarity perspective, compared to a complete integration of large ontologies.

To enable the similarity estimation across several ontologies, once the compared terms are individually linked to concepts of each ontology (by matching their labels), ontologies should be connected in a way that the commonality between the concepts corresponding to the compared terms can be evaluated.

The most basic approaches create an imaginary root node that subsumes the root nodes of different ontologies [11, 13]. Then, the similarity is computed from the resulting knowledge structure. Concretely, in [11], Rodriguez and Egenhofer compute the similarity of terms as the weighted sum of the similarity of three features of their corresponding concepts: the degree of overlapping between synonym sets, conceptual features (e.g. meronyms, attributes, etc.) and neighbour concepts (i.e. those whose path distance to the evaluated concept is lower than or equal to a natural number *r*) [19]. The computation of each of these three components is based on the feature-based similarity measure defined by Tversky [19], which considers the relative importance between common and non-common characteristics. In this approach, two concepts are equivalent if they share the same textual label. The shortest path from each concept to the imaginary root (i.e. depth) is used to give less importance to non-common features, following the idea that individuals pay more attention to similar than to different features during the similarity assessment [19].

Petrakis et al. [13] extended the previous approach relying on the matching between synonym sets and concept glosses extracted from WordNet (i.e. words extracted by parsing concept definitions) or scope notes extracted from MeSH. Once each term to compare is individually matched to a concept of each ontology, they are considered similar if their synonyms (i.e.

different labels corresponding to a single ontological concept) and glosses, and those of the concepts in their neighbourhood are lexically similar. The similarity is computed taking the maximum similarity value obtained by comparing synonyms and glosses per separate using the Jaccard coefficient [25].

Other more recent approaches look for terminologically-equivalent concepts between different ontologies. In Saruladha et al. [1], authors compute semantic similarity among biomedical ontologies based on an information-theoretic perspective of the Tversky's measure [19]. The amount of commonality that exists between concepts c_1 and c_2 is represented by the *IC* of their *LCS (i.e. LCS*(c_1,c_2)), while their differences are conceived as *IC*(c_1) and *IC*(c_2). Considering that both ontologies are connected by a new imaginary root node, the *LCS*(c_1,c_2) is obtained by matching the set of subsumers of c_1 in the first ontology and the set of subsumers of c_2 in the second ontology by means of a terminological matching. The IC of concepts is computed in an intrinsic manner from the knowledge modelled in the ontology, in order to avoid depending on corpora availability [12]. The depth of both concepts in the ontology is computed in the same manner as Rodriguez and Egenhofer [11].

Al-Mubaid and Nguyen [10] also propose a method for assessing similarity of concepts between multiple biomedical ontologies. It compares term pairs using a similarity measure defined in [26], which combines the *path length* and *common specificity* of the corresponding ontological concepts. The common specificity of two concepts is calculated by subtracting the depth of their LCS from the depth of the taxonomic branch to which they belong. Because the path length provides absolute values, the method relies on the selection of a predefined *primary* ontology (the rest are considered as *secondary*) to which similarity values are normalised. A disadvantage of this methodology is the fact that a *primary* ontology must be selected a priori by the user. The differentiation between primary and secondary ontologies makes it necessary to consider a complex casuistic during the similarity assessment. In the multi-ontology scenario, ontologies are connected by joining equivalent taxonomical ancestors (i.e. those with the same textual label). These equivalent concepts are called *bridges*. Because several bridges could be found given a pair of ontologies, evaluated concepts (c_1 and c_2) may have more than one *LCS*, path and common specificity values. Then, if c_1 belongs to the *primary* ontology and c_2 to the *secondary* one, their set of LCS ($LCS_i(c_1,c_2)$) are evaluated considering the *LCS* of the concept that belongs to the primary ontology and the discovered *bridges*: $LCS_i(c_1,c_2) = LCS\ (c_1, bridge_i)$. Then, the maximum similarity obtained considering the full set of *bridges* is taken. When c_1 and c_2 belong to *secondary* ontologies, one of them acts temporarily as primary ontology. Because the similarity measure is based on the path length (which provides absolute values), results obtained from different ontology pairs (with different taxonomical depths and granularity degrees) cannot be compared. Authors propose a method to *scale* the part of the path and the common specificity computed from the *secondary* ontology to the *primary* ontology. The scaling factor is the difference in the taxonomical depth of the *secondary* ontology compared to the *primary* one. Formally, the path length and the common specificity (*CSpec*) are computed as stated in (4) and (5) respectively.

$$Path_i(c_1,c_2) = Path(c_1,bridge_i) + \frac{2 \times D_1 - 1}{2 \times D_2 - 1} \times Path(c_2,bridge_i) - 1 \qquad (4)$$

$$CSpec_i(c_1,c_2) = D_1 - Depth(LCS(c_1,bridge_i)) \qquad (5)$$

where D_1 and D_2 are the depths of the *primary* and *secondary* ontologies respectively.

4 Limitations of related works

The main limitation of the works discussed above is the fact that only terminologically-equivalent ancestors of different ontologies are matched to obtain the LCS. In some cases, it is not possible to find equivalent ancestors for the compared concepts, so that ontologies can only be joined through the root nodes. In other cases, even though a pair of equivalent subsumers is found, another one (more specific), would be more appropriate, but it is omitted because the labels used to refer to subsumers are different (e.g. *cancer/neoplasms*). In both cases, the similarity computed from the terminologically-matched ancestors is lower than the *real* one. In some cases, authors rely on synonyms sets to improve the precision of the matching process.

Terminological matchings are also hampered by the fact that different ontologies very rarely model knowledge in the same way. In fact, as acknowledged by other authors [23], the alignment or merging of different ontologies is a difficult task, because the knowledge representation process is heavily affected by the knowledge engineer point of view, the application in which the ontology will be used, and the distributed nature of the ontology development process.

When trying to discover the LCS of a pair of concepts in different ontologies, we observe that the heterogeneous nature of ontologies hampers related works. On one hand, when modelling large sets of concrete concepts, knowledge engineers progressively group them by introducing common ancestors according to their common characteristics [27]. The level of detail, branching factor and the granularity of the inner taxonomical structure are variables decided by the knowledge expert. Hence, ontologies modelling the same knowledge typically result in very different taxonomical structures. On the other hand, because the knowledge modelling process evolves in a bottom-up fashion, in many situations, ad-hoc abstractions (e.g. *physical entity*, *abstract entity*, *thing* in WordNet) are introduced at the higher levels of the taxonomy. Unlike concrete concepts (e.g. disease names), these ad-hoc abstractions have not been created by consensus and, hence, they are difficult to match.

Example 1. Let us compare the way in which WordNet and MeSH model the concept pair *myocardium* and *heart*, respectively, as shown in Figure 1.

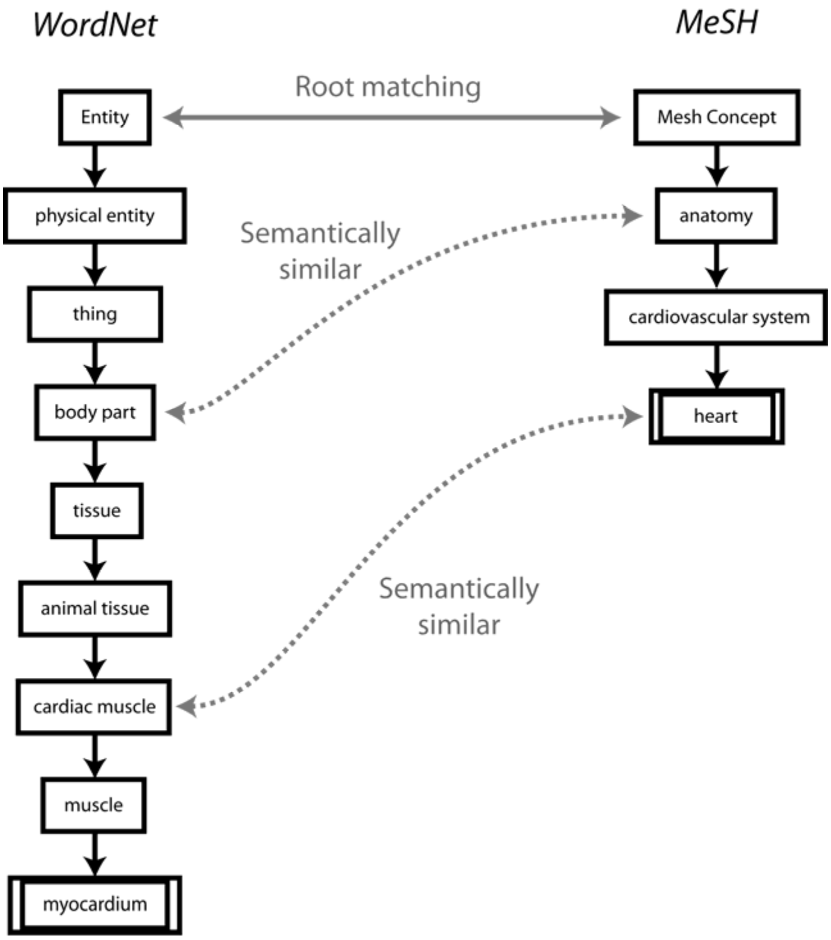

Figure 1. Knowledge modelling for *myocardium* (in WordNet) and *heart* (in MeSH).

Even though both concepts are semantically similar, we observe notorious differences in the granularities of the taxonomical trees. In addition, we realise that subsumers in WordNet tend to present a higher level of abstraction. This is because WordNet has a general scope, whereas MeSH is focused on the biomedical field.

As a result, a terminological matching will not find an equivalent LCS and, hence, both terms will be assessed as maximally distant (i.e. joined by the root nodes). This is the result of comparing heterogeneous taxonomical structures created by knowledge engineers with different points of view. Even though no identical subsumers are available, one realises that some subsumer pairs are semantically similar (e.g. *heart* and *cardiac muscle* or *body part* and *anatomy*) and hence, they could be used as evidence of the sematic commonality between the compared terms (i.e. their LCS).

Example 2. Terms *antibiotic* and *anti-bacterial agent* are evaluated in WordNet and MeSH, respectively, as shown in Figure 2.

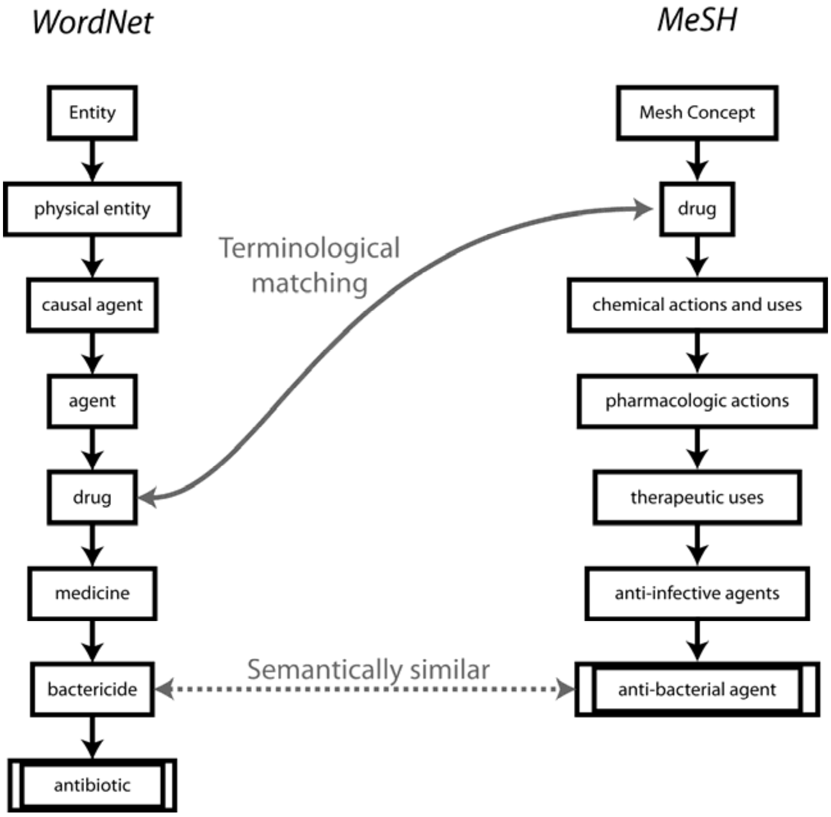

Figure 2. Knowledge modelling for *antibiotic* (in WordNet) and *anti-bacterial agent* (in MeSH).

A terminological matching will be able to select the equivalent subsumer pair *drug* as the LCS. However, as above, another subsumer pair below would be a more adequate LCS (e.g. *bactericide* and *anti-bacterial agents*).

5 Proposed method

In order to overcome the problems discussed in the previous section, our approach aims to discover semantically similar subsumers (but not necessarily terminologically identical), enabling a more accurate assessment of the similarity between the compared terms. Instead of relying solely on the terminological matching between subsumer labels, our approach complements this method by *i)* an assessment of the *semantic overlapping* between subsumer pairs and *ii)* an evaluation of their *structural similarities* analysing the ontologies to which they belong. Both dimensions, considered as explicit and implicit evidences of semantic similarity, are used to quantify the probability of subsumer pairs being equivalent. As a result, two methods for LCS discovery among different ontologies are proposed.

Example 3. To illustrate the definitions and methods formalised in the following subsections, let us consider the ontological structures shown in Figures 3 and 4, associated to the evaluation of the terms *lupus* (in MeSH) and *rheumatoid arthritis* (in WordNet).

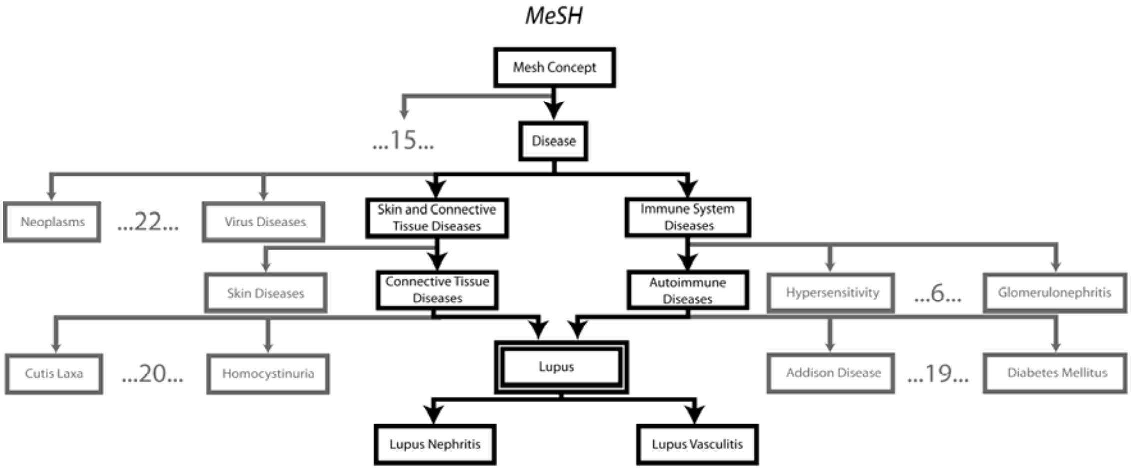

Figure 3. Ontological modelling of the term *lupus* in MeSH.

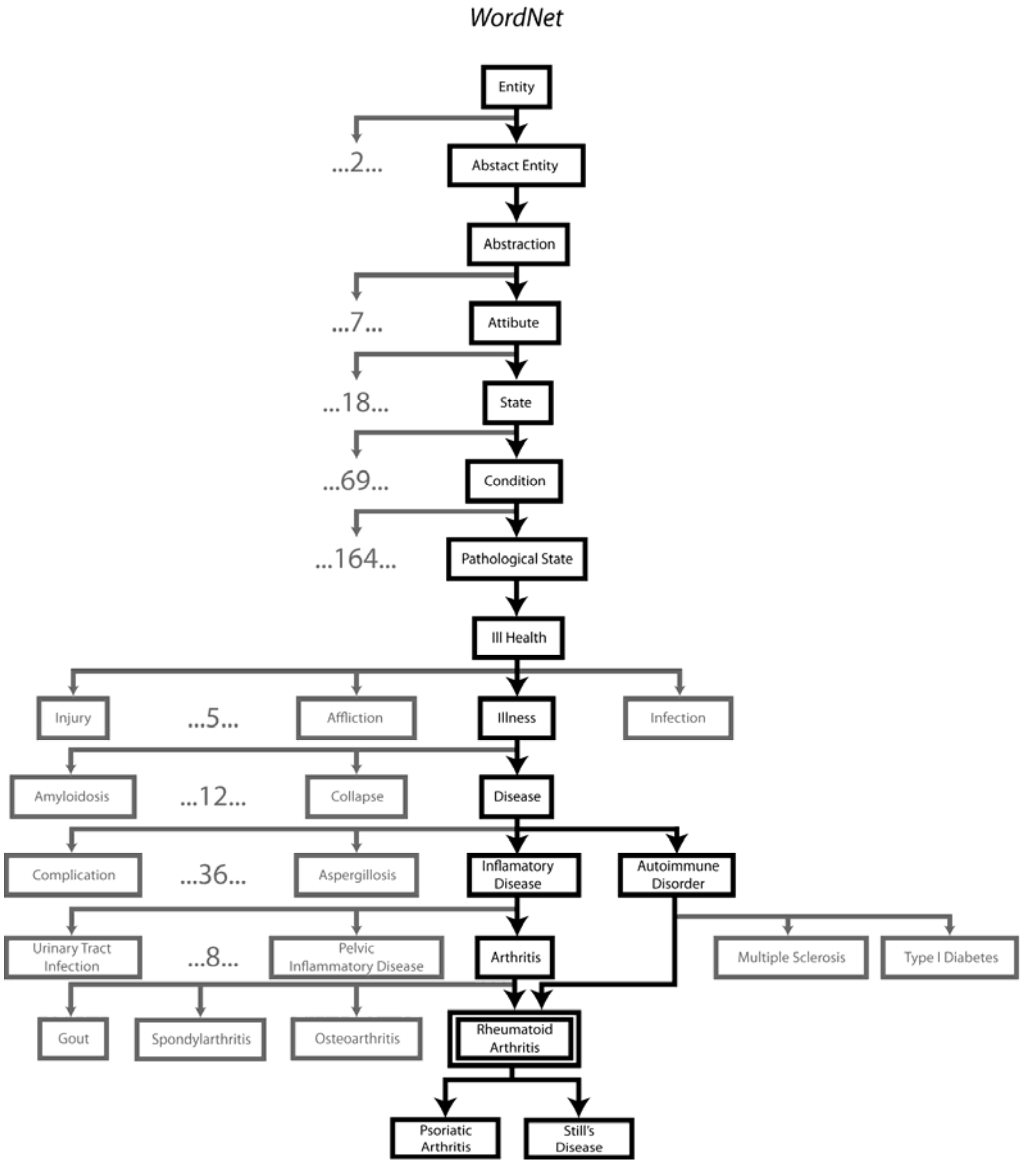

Figure 4. Ontological modelling of the term *rheumatoid arthritis* in WordNet.

5.1 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 interlink distance between a concept and its ancestors. Hence, having two concepts c_i and c_j, $c_i <^n c_j$ implies that c_j is the n^{th} taxonomical ancestor of c_i and, inversely, c_i is the n^{th} specialisation of c_j. Particularly, $c_i <^1 c_j$ is fulfilled if c_i is a *direct* specialisation of c_j; hence, c_j is the *direct* taxonomical ancestor of c_i. Note that $c_i <^0 c_j$ is fulfilled if $c_i = c_j$.

Following Example 3 (Figure 3), *lupus* $<^3$ *disease* is fulfilled because *disease* is the 3rd taxonomical ancestor of *lupus* in MeSH. Likewise, both *connective tissue diseases* and *autoimmune diseases* are a direct taxonomical ancestors of *lupus,* whereas *lupus* is a direct specialization of them (i.e. *lupus* $<^1$ *connective tissue diseases* and *lupus* $<^1$ *autoimmune diseases*).

Definition 2. The *closure* of this relation (<*) is the union of the result of applying $<^n$ for n=0… d, being d the maximum number taxonomical links between a concept and the *root* node. Consequently, $c_i <^* c_j$ is fulfilled if c_j is an ancestor of c_i at *any* level in the taxonomical tree or if $c_i = c_j$; inversely, for a given c_j, c_i is a taxonomical specialisation of c_j at any taxonomical depth.

Applied to Example 3 (Figure 3), the closure of the subsumption relation (<*) for *lupus* in MeSH is fulfilled for the set *{lupus, connective tissue diseases, autoimmune diseases, skin and connective tissue diseases, immune system diseases, diseases and MeSH concept}* that covers all the taxonomical ancestors (considering multiple taxonomical inheritance) of *lupus* including itself.

Definition 3. The set of *subsumers* of a concept c in the ontology O at distance n is:

$$subsumers_O^n(c) = \{s \in C \mid c <^n s\}$$

Definition 4. The set of *direct subsumers* of a concept c in the ontology O is:

$$direct\_sub_O(c) = subsumers_O^1(c) = \{s \in C \mid c <^1 s\}$$

Definition 5. The complete set of *subsumers* of a concept *c* in the ontology *O* at any taxonomical level is:

$$total\_sub_O(c) = subsumers_O^*(c) = \{s \in C \mid c <^* s\}$$

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

In Example 3 (Figure 3): *direct_sub$_{MeSH}$(lupus)={connective tissue diseases, autoimmune diseases}*, *total_sub$_{MeSH}$(lupus)= { lupus, connective tissue diseases, autoimmune diseases, skin and connective tissue diseases, immune system diseases, diseases and MeSH concept }.*

Definition 6. The set of *hyponyms* of a concept *c* in the ontology *O* at distance *n* is:

$$hyponyms_O^n(c) = \{h \in C \mid h <^n c\}$$

Definition 7. The set of *direct hyponyms* of a concept *c* in the ontology *O* is:

$$direct\_hypo_O(c) = hyponyms_O^1(c) = \{h \in C \mid h <^1 c\}$$

Definition 8. The complete set of *hyponyms* of a concept *c* in the ontology *O* at any taxonomical depth is:

$$total\_hypo_O(c) = hyponyms_O^*(c) = \{h \in C \mid h <^* c\}$$

In Example 3 (Figure 4), considering the *arthritis* concept in WordNet:

direct_hypo$_{WordNet}$(arthritis)={ rheumatoid arthritis, osteoarthritis, spondylarthritis, gout},

total_hypo$_{WordNet}$(arthritis)={arthritis, rheumatoid arthritis, osteoarthritis, spondylarthritis, gout, psoriatic arthritis, Still's disease}.

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

5.2 Analysing the semantic overlapping between subsumers

The first aspect considered when comparing subsumer pairs of different ontologies is their degree of semantic overlapping. Similarly to approaches that compute the IC of a concept from an ontology in an intrinsic manner [17, 28] (introduced in section 2), we assess the semantic content of a concept according to its taxonomical specialisations (i.e. hyponyms). This is based on the principle of cognitive saliency [29]: concepts are specialised when they must be differentiated from other ones. Hence, the set hyponyms of a concept summarises and bounds its meaning, differentiating it from other concepts. For example, the meaning of the concept *body part* is the result of the sum of all its specialisations (i.e. anatomical entities). Interpreting this principle in an inverse manner, the fact that two subsumers (each one modelled in a different ontology) share a certain amount of hyponyms, gives us an evidence of semantic similarity. As a result, comparing the set of hyponyms of two subsumers of different ontologies, we are able to quantify their semantic overlapping and, hence, measure the degree of semantic equivalence of two subsumers.

More in detail, given a pair of concepts c_1, c_2 to be evaluated, where c_1 belongs to the ontology O_1 and c_2 belongs to O_2, their subsumer pairs ($<s_i,s_j>$ where $s_i \in total\_sub_{O_1}(c_1)$ and $s_j \in total\_sub_{O_2}(c_2)$) can be compared according to the degree of overlapping of their hyponym sets ($total\_hypo_{O_1}(s_i)$ and $total\_hypo_{O_2}(s_j)$).

Hyponym sets are compared by means of terminological matchings between their labels. The matching of hyponyms, however, is more effective than that of subsumers, thanks to the higher dimensionality of hyponym sets (especially when dealing with abstract subsumers), and due to the fact that hyponyms refer to more concrete concepts, being less ambiguous, less affected by synonymy and usually labelled with consensus words (e.g. concrete disease names). It is important to note that, even though the terminological matching between hyponyms sets underestimates the *real* semantic commonalities between subsumer pairs, all of them will be evaluated in the same way, enabling an objective quantification of the most similar pair.

After applying this terminological matching, the degree of overlapping between the sets of hyponyms is quantified using a similarity coefficient. Many coefficients have been proposed to evaluate set representations [25, 30-33] (some of the most commonly used ones are shown in Table I). All of them quantify the level of commonality between sets (i.e. intersection) whereas their main difference is the normalising factor. Jaccard's coefficient divides the intersection by union of the compared sets. The coefficients of Dice, Ochiai, Simpson and Braun-Blanquet use an averaging operator to weight the contribution of both sets (arithmetic mean, geometric mean, minimum and maximum size, respectively). In our approach, we opted for the Ochiai coefficient due to the geometric mean, while considering the size of both sets, tends to be lower than the arithmetic or absolute sum when one of the sets is small. This results in higher similarity values and, hence, in a prioritisation of the commonalities of subsumers at a lower level of the hierarchy, which, obviously, will present less hyponyms than those at a higher level. This strategy implicitly considers the relative depth of subsumers in the hierarchy (as a function of the hyponym set size), a dimension that, as discussed in section 2, is desirable for similarity assessment.

Table I. Similarity coefficients, where A and B are the compared sets.

Coefficient	*Equation*
Jaccard [25]	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Dice [32]	$\frac{2 \times \lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$
Ochiai [30]	$\frac{\lvert A \cap B \rvert}{\sqrt{\lvert A \rvert \times \lvert B \rvert}}$
Simpson [31]	$\frac{\lvert A \cap B \rvert}{Min(\lvert A \rvert, \lvert B \rvert)}$
Braun-Blanquet [33]	$\frac{\lvert A \cap B \rvert}{Max(\lvert A \rvert, \lvert B \rvert)}$

Formally, being s_i, s_j two subsumers and being O_1 and O_2 two ontologies so that s_i belongs to O_1 and s_j belongs to O_2 the *semantic overlapping* between them is computed as:

$$sem\_overlap(s_i, s_j) = \frac{\lvert total\_hypo_{O_1}(s_i) \cap total\_hypo_{O_2}(s_j) \rvert}{\sqrt{\lvert total\_hypo_{O_1}(s_i) \rvert \times \lvert total\_hypo_{O_2}(s_j) \rvert}} \qquad (6)$$

where the intersection ($\cap$) between both sets of hyponyms is defined as the set of concepts that are terminologically-equivalent (i.e. their labels or the labels of their synonyms, if available, are identical).

5.3 Evaluating the structural similarity between subsumers

Considering that ontological structures provide implicit evidences concept semantics [4], in this section, we present a measure that, inspired by graph-matching theory, aims to quantify the structural similarities between concept subsumers of different ontologies. Coherent with the knowledge representation principles, the measure assumes that if two subsumers are equivalent their related ancestors and hyponyms should be semantically similar. Analogously to section

5.2, this measure will contribute to the discovery of more suitable subsumers for the compared concepts.

Relying on the graph theory, the knowledge structure defined by the semantic relationships modelled in ontologies can be represented by means of attributed graphs, where nodes of the graph are concepts and relations between nodes are semantic relations (in our case, focused on taxonomic links). Using a graph representation, graph matching algorithms can be applied to find equivalences between ontologies according to their structural similarities. These algorithms aim to find a bijection between nodes of two graphs considering nodes (i.e. concepts) and relations jointly. A large number of solutions exist [34-36], some of them being specifically applied to the ontology matching problem [37, 38].

The discovery of bijections between graphs is closely related to the discovery of the LCS of a pair of concepts among different ontologies. Inspired by the graph matching methods proposed in [34] or in [39], which were applied to attributed graphs, we propose evaluating the structural similarity between subsumer pairs according to the *degree of matching* of their adjacent nodes (i.e. those that related them with their *direct subsumers* and *direct hyponyms*, as formalised in Definition 4 and 7). Additional levels of adjacency could be considered to potentially obtain a more precise assessment (e.g. those edges ending at $hyponyms_O^2(s)$ and $subsumers_O^2(s)$). However, the computational complexity to evaluate how similar two subsumers are increases in a non-polynomial way with respect the levels of adjacency considered. Consequently, for scalability reasons, and considering the size of widely-used ontologies like WordNet or those in the UMLS repository, we have limited the analysis to direct incident edges.

It is important to note that ingoing relations (i.e. those relating a subsumer with its *direct subsumers*) and outgoing relations (i.e. those relating a subsumer with its *direct hyponyms*) are considered separately. This differentiation is mandatory because directionality of edges must be preserved to properly evaluate the modelled semantics.

The degree of matching of a relation pair (each one of a different subsumer, s_i, s_j) is measured as a function of the similarity between the nodes at the end of the relation (i.e. *direct_sub(s_i)* vs *direct_sub(s_j)* and *direct_hypo(s_i)* vs *direct_hypo(s_j)*). In order to provide a semantically-grounded assessment, this is measured according to degree of *semantic overlapping* (eq. 6) between nodes (i.e. concepts), as proposed in section 5.2.

When comparing outgoing or ingoing relation sets of a subsumer s_i (in an ontology) to those of another subsumer s_j (in a different ontology), we do not rely on the cardinality of these sets (e.g. the fact that the first subsumer has 3 direct hyponyms whereas the other one has only 2). We believe that the consideration of cardinalities is inconsistent due to variations in the levels of detail of different ontologies, which result in different granularities and branching factors of taxonomical structures (even modelling the same knowledge). As discussed in section 4, this is again related to the fact of comparing heterogeneously constructed sources, and due to the subjectivity inherent to manual knowledge modelling. On the contrary, in our proposal, the degree of matching of each edge of s_i to those of s_j is computed as the *maximum degree of matching* to *any* edge of s_j (considering the directionality of the relations).

To illustrate this process, let us evaluate the structural similarities for the term *rheumatoid arthritis* (in WordNet) and *lupus* (in MeSH) as presented in Example 3. Figure 5 shows an extract of the adjacent ontological structures for both concepts.

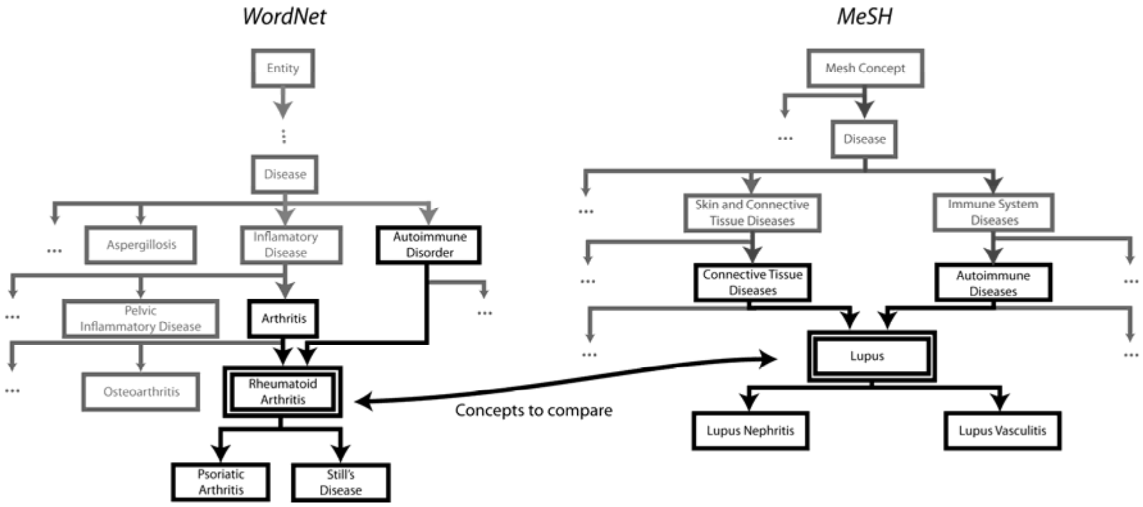

Figure 5. Structural evaluation of *rheumatoid arthritis* (in WordNet) and *lupus* (in MeSH).

To evaluate the structural similarity of both concepts, the hyponymy (i.e. outgoing) relations of *rheumatoid arthritis* with *psoriatic arthritis* and *Still's disease* in WordNet will be compared to relations of *lupus* with *lupus nephritis* and *lupus vasculitis* in MeSH. On the other hand, generalisation (i.e. ingoing) relations of *rheumatoid arthritis* with *arthritis* and *autoimmune disorder* in WordNet will be compared to ingoing relations of *lupus* in MeSH, in this case, *connective tissue disease* and *autoimmune diseases*. The degree of matching of each relation of an ontology (e.g. *rheumatoid arthritis- autoimmune disorder* in WordNet) with the relations in the other ontology (i.e. *lupus-connective tissue diseases* and *lupus-autoimmune diseases* in MeSH) will be assessed as the *maximum degree of matching* of the former to any of the latter. As stated above, this is computed as the degree of *semantic overlapping* between nodes at the end of the compared relations. Hence, when evaluating the relation (*rheumatoid arthritis-autoimmune disorder*) of WordNet, the semantic overlapping between concept pairs (*autoimmune disorder, connective tissue diseases*) and (*autoimmune disorder*, *autoimmune diseases*) will be computed; the maximum value is taken as the degree of matching of the relation (*rheumatoid arthritis-autoimmune disorder*).

After that, when all relation pairs (both ingoing and outgoing) have been assessed, the *structural similarity* for s_i compared to s_j is computed as the *average* of the degree of matching of the relations of s_i in O_1 to the relations of s_j in O_2. Formally:

$$structural\_similarity_{s_j}(s_i) = \frac{\sum\limits_{s_{ir} \in direct\_sub(s_i)} \max\limits_{s_{jp} \in direct\_sub(s_j)} \{sem\_overlap(s_{ir}, s_{jp})\} + \\ + \sum\limits_{h_{ir} \in direct\_hypo(s_i)} \max\limits_{h_{jp} \in direct\_hypo(s_j)} \{sem\_overlap(h_{ir}, h_{jp})\}}{|\, direct\_sub(s_i) \,| + |\, direct\_hypo(s_i) \,|} \quad (7)$$

Note that the structural similarity of s_i in O_1 compared to s_j in O_2 could be different to the similarity of s_j in O_2 compared to s_i in O_1, due to divergences in relation cardinalities (i.e. it is not a symmetric function). Hence, we define the *pairwise structural similarity* for a pair of subsumers as the maximum of the individual similarities computed from O_1 to O_2 and from O_2 to O_1. Formally:

$$pairwise\_struc\_sim(s_i, s_j) = \max(structural\_similarity_{s_j}(s_i), structural\_similarity_{s_i}(s_j)) \quad (8)$$

5.4 Selecting the LCS

Once we are able to quantify both the semantic (i.e. explicit) and structural (i.e. implicit) similarity between subsumer pairs of different ontologies, we can select the most suitable one as the LCS. In this section, we present two methods to discover it: the first one, based solely on semantic overlapping, and the second one, which also considers structural similarity. In the following, the common steps to both methods are detailed. After that, the different strategies implemented by each method are detailed in sections 5.4.1 and 5.4.2 respectively.

First, both methods apply the terminological matching used by related works [1, 10]. Subsumers of the compared concepts (each one belonging to a different ontology) with identical labels are matched. If no terminologically-equivalent pairs are found, the root nodes of both ontologies are matched; otherwise, if one or more pairs are identical, the most concrete pair of equivalent ancestors is selected. We consider the most concrete pair of subsumers is the one that minimises the path length between the compared terms. As a result a pair of terminologically-matched subsumers (each one belonging to a different ontology) is obtained. Formally:

Definition 9. Given a pair of concepts c_1, c_2 so that c_1 belongs to O_1 and c_2 belongs to O_2, and given $total\_sub_{O1}(c_1)$, $total\_sub_{O2}(c_2)$, their pair of terminologically-matched subsumers ($<ms_1, ms_2>$) is:

$$< ms_1, ms_2 >= \begin{cases} \arg\min\limits_{\forall <s_i, s_j>} (path(c_1, c_2)) \mid < s_i \in total\_sub_{O_1}(c_1), s_j \in total\_sub_{O_2}(c_2) > & , if\ s_i = s_j \\ < root\_node(O_1), root\_node(O_2) > & , otherwise \end{cases}$$

where the '=' operator considers identical labels (including synonyms sets).

In Example 3, when comparing *lupus* (in MeSH) and *rheumatoid arthritis* (in WordNet), we have $total\_sub_{MeSH}$*(lupus)={lupus, connective tissue diseases, autoimmune diseases, skin and connective tissue diseases, immune system diseases, disease, MeSH concept}* and $total\_sub_{WordNet}$*(rheumatoid arthritis)={rheumatoid arthritis, arthritis, inflammatory disease, disease, illness, ill health, pathological state, condition, state, attribute, abstraction, abstract entity, entity}*. The most specific subsumer pair $<ms_1, ms_2>$ so that ms_1 belongs to $total\_sub_{MeSH}$*(lupus)* and ms_2 belongs to $total\_sub_{WordNet}$*(rheumatoid arthritis)* that match terminologically is <*disease, disease*>.

Then, by means of the semantic and structural measures proposed in section 5.2 and 5.3, we are able to assess the degree of similarity of the subsumer pair (ms_1, ms_2). This value is taken as baseline to find a more adequate subsumer pair below them (in their respective taxonomical trees) and above the compared concepts with a higher resemblance. As discussed in section 4, this indicates the presence of other subsumer pairs that, even though they have no identical labels, are better suited to act as the LCS. Formally:

Definition 10. For each c_i in $\{c_1,c_2\}$, the subset of subsumers to be semantically and/or structurally compared in order to select a more suitable LCS are those below their matched subsumer pair (ms_1, ms_2) and each concept:

$$candidate\_LCS_{O_i}(c_i) = \{s \in C_i \mid c_i <^* s \wedge s_i <^* ms_i\}$$

Note that, according to the subsumption relation (Definition 1), both the concept itself (c_i) and the matched subsumer (ms_i) are contained in the set and, hence, are candidates for being LCS.

In Example 3, the *candidate_LCS*$_{MeSH}$*(lupus)={lupus, connective tissue diseases, autoimmune diseases, skin and connective tissue diseases, immune diseases, disease}*, whereas the *candidate_LCS*$_{WordNet}$*(rheumatoid arthritis)={rheumatoid arthritis, arthritis, inflammatory disease, autoimmune disorder, disease}.*

5.4.1 Selecting the LCS considering semantic information

The first method considers the explicit degree of semantic overlapping between subsumer pairs to select the LCS. The pair with the maximum overlapping is taken as the final LCS:

$$LCS(c_1, c_2) = \underset{\forall < cs_{1i}, cs_{2j} >}{\arg\max} \{sem\_overlap(cs_{1i}, cs_{2j})\} \quad (9)$$

where<cs_{1i},cs_{2j}> is any tuple of subsumers resulting from the Cartesian product between the sets of candidate LCS:

$$< cs_{1i}, cs_{2j} > \in \{candidate\_LCS_{O_1}(c_1) \times candidate\_LCS_{O_2}(c_2)\} \quad (10)$$

Because only subsumer sets are evaluated in this method (around a dozen of elements in large ontologies like WordNet), its computational complexity is low, resulting in a highly scalable method for large dataset and ontologies.

5.4.2 Selecting the LCS considering structural and semantic information

The second method complements the semantic overlapping with the structural similarity. Again, the Cartesian product between *candidate LCS* is evaluated. Each subsumer pair is compared both from a structural and semantic perspective (eq. (6) and (8), respectively). The *average* of both scores is taken (11):

$$subsumer\_similarity(cs_{1i}, cs_{2j}) = \frac{pairwise\_struc\_sim(cs_{1i}, cs_{2j}) + sem\_overlap(cs_{1i}, cs_{2j})}{2} \qquad (11)$$

Then, in order to prioritise subsumer pairs located at a lower level of the taxonomical tree (which, due to their higher level of specialisation tend to be less semantically distant, as discussed in section 2), we normalise the value by the path length resulting from going from c_1 to c_2 through the evaluated subsumer pair (cs_{1i}, cs_{2j}). Again, this implicitly considers the taxonomical depth of concepts as an important dimension of the semantic assessment.

As in section 5.4.1, the final LCS is the pair which maximises the resulting score (12):

$$LCS(c_1, c_2) = \underset{\forall < cs_{1i}, cs_{2j} >}{\arg\max} \left\{ \frac{subsumer\_similarity(cs_{1i}, cs_{2j})}{N_1 + N_2} \right\} \qquad (12)$$

where N_1 and N_2 are the minimum number of taxonomical links from c_1 to cs_i and from c_2 to cs_j, respectively.

As it is shown in the results section, the evaluation of both semantic and structural features improves, in most cases, the LCS selection and hence the similarity assessments. This method, however, requires the evaluation of a higher number of ontological concepts (i.e. subsumers and concepts in their direct adjacency) than the first method.

6 Results

The evaluation of similarity measures is usually performed by comparing (i.e. computing the degree of correlation) the automatically obtained similarity values with those provided by human experts [9, 10, 13]. To allow the reproducibility of the evaluation experiments, several authors proposed different benchmarks [40, 41] consisting of word pairs whose similarity have been evaluated by a group of experts.

In the biomedical field, we can find the benchmark of Pedersen et al. [9] (by far, the most used one [18]) and the one by Hliaoutakis et al. [42]. The first one consists of 30 pairs of medical terms whose similarity was assessed by experts of the Mayo Clinic. A total of 3 physicians and 9 medical coders evaluated each word pair. After a normalisation process, the average similarity values provided by both sets of experts in a scale from 1 (non-similar) to 4 (identical) were obtained. The second benchmark is composed of a set of 36 medical terms extracted from the MeSH repository. The similarity between each pair was assessed by 8 medical experts from 0 (non-similar) to 1 (identical).

Unfortunately, standard evaluation benchmarks focused on multi-ontology similarity methods have not been proposed [13]. Related works commonly use two considerably different ontologies (e.g. WordNet and MeSH), take some of the benchmarks of word pairs mentioned above and assess their similarity considering that each of the two terms of each pair belongs to an unique ontology (ignoring the fact that it could be found in both ontologies) [13, 43]; results are compared to similarity ratings provided by human experts.

We carried out an experiment similar to the one proposed in [13]. Two ontologies are used as background knowledge: WordNet and MeSH. WordNet [7] is a lexical database that describes and structures more than 100,000 general concepts, which are semantically structured in an ontological way. WordNet contains English words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets), each one expressing a distinct concept. Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (subclass-of), meronymy (part-of), etc. We used WordNet version 2 in our tests as it is the most common version used in the related works. The Medical Subject Headings (MeSH) [44] contains a hierarchy of medical and biological terms defined by the U.S National Library of Medicine. This classification was initially created to catalogue books and other library materials, and to index articles for inclusion in health-related databases including MEDLINE. In MeSH tree, there are 16 basic categories, with more than 22,000 concepts. We used the latest 2011 MeSH XML files available for download[i].

The reason for combining a general-purpose ontology like WordNet with a domain-specific one, such as MeSH, (as done in [13]) is to consider a least favourable evaluation scenario. Recall that both ontologies have been designed with significantly different scopes. Hence, the modelled subsumers and taxonomical trees tend to be significantly different for a pair of similar terms (some examples are given in section 4). On one hand, the differences in scope are reflected by the fact that only 5.4% of WordNet concepts appear in MeSH with the same textual label (including synonyms). This will affect the accuracy of methods based solely on terminological matchings, showing the advantages of a semantically-grounded method.

[i] http://www.nlm.nih.gov/mesh/filelist.html

We use the benchmarks of Hliaoutakis et al. and Pedersen et al. to evaluate our methods. For the first one, we have taken the ratings given by the 3 physicians, 9 medical coders and the average of both. Due to the fact that these benchmarks are meant to evaluate mono-ontology similarity measures, most of the word pairs can be found both in MeSH and in WordNet. In fact, 20 out of the 30 word pairs of the benchmark of Pedersen et al. and 35 out of 36 word pairs of the Hliaoutakis et al. benchmark can be found in both ontologies (see tables II and III). On one hand, we considered only those pairs that can be found in both ontologies in order to compare the similarity accuracy obtained when using a unique (and semantically coherent) knowledge source with the more challenging multi-ontology scenario. In fact, the mono-ontology similarity accuracies give us an idea about the ground truth precision that one can expect from the knowledge sources, and helps to contextualise the results when both ontologies are combined. On the other hand, to evaluate the multi-ontology scenario, one of the two terms of each pair is considered to be in WordNet (regardless it can also be found in MeSH), whereas the second term is considered to be in MeSH (regardless being found in WordNet), as done in [13]. In this manner, a suitable LCS between the two ontologies must be discovered to enable the similarity assessment. Tables II and III show which terms of each benchmark were evaluated in which ontology.

Table II. Medical term pairs with averaged similarity scores of experts extracted from the benchmark of Pedersen et al. [9].

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

Table III. Medical term pairs with averaged similarity scores of experts extracted from the Hliaoutakis et al. benchmark [42].

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

In addition to compare the accuracy of our methods with the (most favourable) mono-ontology setting, we also compared them against the strategies used by related works in the multi-ontology scenario. To do so, four strategies to select the LCS have been implemented for the multi-ontology scenario:

- *S1*: The root nodes of the two ontologies are joined and considered as the LCS, following the strategy proposed in [11, 13].
- *S2*: The most specific (i.e. deepest) pair of subsumers for the compared concepts that are terminologically-equivalent (considering synonyms) among the two ontologies are taken as the LCS, as in [1, 10].
- *S3*: The LCS is selected based on our first method proposed in section 5.4.1, relaying on the semantic knowledge modelled in the taxonomy.
- *S4*: The LCS is selected based on our second method proposed in section 5.4.2, also considering additional structural evidences.

For all configurations and benchmarks, as similarity measure, we have used the three path-based functions introduced in section 2: Rada (eq. (1)), Wu & Palmer (W&P) (eq. (2)) and Leacock and Chodorow (L&C) (eq. (3)). Note that, when using the Wu & Palmer measure in the multi-ontology setting, the depth of the LCS may be different from one ontology to another (see some examples in section 4). Following the same premise of edge-counting measures, which always evaluate the shortest path (i.e. the one that gives the highest evidence of similarity), we have taken the minimum depth value during the assessments.

Correlation values against human judgments for the different methods, measures, benchmarks and ontology combinations are summarised in table IV.

Table IV. Correlation values for three edge-counting measures (Rada, Wu & Palmer (W&P) and Leacock and Chodorow (L&C)) against the set of 20 word pairs from the benchmark of Pedersen et al. (ratings of the 3 physicians, ratings, of the 9 medical coders and the average of all of them) and the set of 35 word pairs from the Hliaoutakis et al. benchmark in a mono ontology scenario (MeSH or WordNet only), and in a multi-ontology setting (MeSH+WordNet) varying strategy to select the LCS (S1, S2, S3 and S4, **bold** lines represent our methods).

Ontologies	Similarity measure	LCS selection strategy	Pedersen et al. Physicians	Pedersen et al. Coders	Pedersen et al. All (averaged)	Hliaoutakis et al.
MeSH	Rada	-	0.66	0.63	0.67	0.68
WordNet	Rada	-	0.39	0.48	0.45	0.53
MeSH + WordNet	Rada	S1	-0.16	0	-0.08	-0.21
MeSH + WordNet	Rada	S2	0.35	0.19	0.27	0.63
MeSH + WordNet	**Rada**	**S3**	**0.48**	**0.55**	**0.53**	**0.65**
MeSH + WordNet	**Rada**	**S4**	**0.59**	**0.64**	**0.64**	**0.67**
MeSH	W&P	-	0.66	0.66	0.68	0.69
WordNet	W&P	-	0.40	0.43	0.43	0.53
MeSH + WordNet	W&P	S1	-0.17	-0.04	-0.10	-0.16
MeSH + WordNet	W&P	S2	0.41	0.29	0.36	0.67
MeSH + WordNet	**W&P**	**S3**	**0.51**	**0.62**	**0.59**	**0.67**
MeSH + WordNet	**W&P**	**S4**	**0.58**	**0.74**	**0.68**	**0.67**
MeSH	L&C	-	0.67	0.74	0.73	0.74
WordNet	L&C	-	0.53	0.66	0.62	0.66
MeSH + WordNet	L&C	S1	-0.17	-0.02	-0.09	-0.18
MeSH + WordNet	L&C	S2	0.44	0.36	0.41	0.68
MeSH + WordNet	**L&C**	**S3**	**0.58**	**0.73**	**0.68**	**0.71**
MeSH + WordNet	**L&C**	**S4**	**0.66**	**0.77**	**0.74**	**0.72**

7 Discussion

Analysing the results shown in table IV, several conclusions may be drawn. First, one notices differences in correlation values for the two benchmarks when evaluating word pairs in MeSH and in WordNet individually. Similarity values assessed from WordNet tend to be significantly lower than those from MeSH for all the benchmarks and measures (e.g. 0.63-67 vs. 0.39-48 for the Pedersen et al. benchmark and the Rada measure). This illustrates the differences in knowledge modelling between the two ontologies for the same concepts, and how MeSH provides a more accurate knowledge representation of medical concepts. However, it is worth noting that the differences for the Hliaoutakis et al. benchmark are lower than for the benchmark of Pedersen et al. because the former contains more common medical terms (in fact, most terms in the Hliaoutakis et al. benchmark were contained in WordNet, as detailed in section 6).

Ideally, in the multi-ontology scenario (in which a term of each pair is evaluated in WordNet, the other term in MeSH and a common LCS is discovered), one would expect a correlation value in a range between the correlations provided when using the two ontologies individually. The closer to the MeSH correlation (which is the highest in the mono-ontology setting), the better the assessment of the LCS in the multi-ontology setting would be. Analysing the results obtained in the multi-ontology scenario for the benchmark of Pedersen et al., we observe very low correlations when relying on non-semantic matchings between subsumers. On one hand, the naïve approach based on joining root nodes (S1) provides completely uncorrelated results (very near to zero or even below zero). With this strategy, all pairs appear to be maximally distant and, hence, hardly distinguishable. On the other hand, the method based on matching terminologically-equivalent subsumers (S2) depends heavily on the labels used to refer to those subsumers. Thanks to the degree of terminological overlapping between WordNet and MeSH, this strategy is able to improve the naïve method, even though correlation values are still lower than the worst mono-ontology assessment (e.g. between 0.19 and 0.35 for the Rada measure). Analysing the results provided by our methods (S3 and S4) for the same tests, we observe a clear improvement. By considering only the semantic overlapping (as presented in section 5.4.1), we are able to increase the correlation to values that are in the range defined by the mono-ontology results (e.g. correlation values of 0.48-0.55, which are in the range 0.39…0.67, obtained for the Rada measure in a mono-ontology setting). Evaluating the explicit semantics provided by ontologies (i.e. hyponym sets), our first method (S3) is able to match more suitable subsumers than the terminological matching. When we also take into consideration structural similarities between subsumers (S4, as proposed in section 5.4.2), correlation values increase, being very close (or even higher in some cases) to the best mono-ontology result (e.g. 0.59-0.64 in the range 0.39…0.67, obtained for the Rada measure in a mono-ontology setting). By considering the similarity between concepts in the immediate adjacency of each subsumer pair, normalised by the path length resulting from going through them, we are able to strengthen the confidence on those subsumer pairs that maximise the similarity between the compared concepts. Hence, as we are capturing more explicit and implicit knowledge, we are able to better quantify their commonalities. Obviously, our second method (S4) requires the analysis of a larger set of ontological concepts (around a few hundreds in these tests) than our first one (S1) (around a few dozens in these tests), which provides a high scalability at the expense of lower accuracy.

Differences between correlation values for ratings provided by coders and physicians for the Pedersen et al. benchmark are also worth noting. As stated by the authors [9], during the construction of the benchmark, medical coders were asked to reproduce the classic Rubenstein and Goodenough [41] and Miller and Charles [40] tests to ensure that they understood the notion of similarity. Physicians, however, rated concept pairs without pre-training. As a consequence, medical coders, who were trained and more used to the definition of hierarchical classifications, provided ratings that seem to better reproduce the concept of *taxonomic similarity* [18]. Therefore, because our methods rely on taxonomical evidences (both semantic and structural), it is coherent that our results correlate better with coders than with physicians.

Results obtained for the Hliaoutakis et al. benchmark are less illustrative. We observe that considering only terminological matchings (S2), correlation values are surprisingly high, and very close to the best mono-ontology scenario (e.g. 0.63 vs. 0.68 with the Rada measure). This means that, for most of the word pairs, it was possible to discover a terminologically-equivalent subsumer pair that best represents the commonalities between them. As a result, considering the correlation range for the mono-ontology scenario (e.g. 0.53 for WordNet and 0.68 for MeSH using the Rada measure), very low improvement is possible when using a semantic approach. In any case, the results provided by our methods (S3 and S4) are coherent to what was observed for the benchmark of Pedersen et al. , providing a slight improvement as more evidences (semantic and/or structural) are considered (e.g. 0.65 for S3 and 0.67 for S4, using the Rada measure).

Differences between each similarity measure are also coherent to what was commented above. The ground truth correlations tend to be higher for the Wu & Palmer (W&P) and Leacock & Chodorow (L&C) measures than for the more basic Rada measure. As stated in section 2, the evaluation of the taxonomical depth in similarity assessment tends to improve the results because it helps to distinguish concept pairs with different levels of abstraction. Moreover, it also acts as a normalising factor for the path length. This is very convenient in the multi-ontology setting because it enables direct comparisons among similarity values computed from different ontology pairs. The absolute path length value, on the contrary, cannot be directly compared among different ontologies, as discussed in [10]. The improvement for these measures is particularly noticeable for coders' ratings when applying the second method (S4) (0.74 for Wu & Palmer and 0.77 for Leacock & Chodorow), offering accuracies that are even higher than those obtained for the mono-ontology setting (0.66 and 0.74 in the best case, respectively). This shows how a proper integration of heterogeneous sources may overcome the shortcomings of individual ontologies (i.e. the fact that an individual concept is more accurately modelled in an ontology than in another one).

8 Conclusions

Semantic similarity estimation contributes to the better understanding of textual resources. It has been successfully applied in many areas such as word sense disambiguation [45], synonym detection [16], automatic spelling error detection and correction [46], automatic language translation [47], information extraction [48, 49], document categorisation or clustering [47], semantic annotation [50] and ontology learning [51-53]. In the biomedical field, due to the importance of information, much of which is presented in a textual form, semantic similarity measures have been of great interest [54]. These measures help to classify medical data [55, 56], organise medical literature [57], integrate of heterogeneous clinical data [58], or improve information retrieval tasks [9, 59].

Among the different approaches proposed for similarity estimation, ontology-based ones have proven to be one of the most effective [8, 17, 18]. Most of these approaches, however, are limited by the fact that a unique ontology is exploited. Hence, they rely completely on the coverage and completeness of the input ontology. The exploitation of multiple ontologies for similarity assessment may be necessary, for example, in semantic information retrieval from biomedical sources, in which the similarity between queries and document terms should be assessed [1, 9, 10]. On one hand, medical ontologies (e.g. those in the UMLS) are created with different scopes and purposes (e.g. MeSH aims to index and catalogue medical literature, whereas SNOMED CT is meant to bring structure to diseases, clinical findings, procedures, etc.), offering partial views of the same domain. On the other hand, terms contained in documents or queries may refer to general concepts that can only be found in general repositories such as WordNet. In both cases, it is needed to assess the semantic similarity across different ontologies. In fact, the ability to detect similar concepts (e.g. identical concepts referred with synonym terms) across different ontologies is also interesting when aiming to exploit or even integrate heterogeneous data sources (e.g. electronic health care records, medical databases, etc.). In the context of semantic information retrieval, this can be useful to detect or propose equivalent query formulations that can be useful to improve the retrieval recall in tasks, such as patient cohort identification [9].

In this paper, a general approach to enable the similarity assessment across multiple ontologies is presented. Because most ontology-based measures focus the similarity estimation on the LCS of the compared terms, our methods (the first one focused on high scalability and the second one centred on high accuracy) aim to discover a LCS among several ontologies that accurately represents the commonalities between terms. Our approach is based on explicit (semantic) and implicit (structural) evidences observed in the background ontologies. The evaluation, based on well-known benchmarks of biomedical terms and widely used ontologies, has shown an increase in the similarity accuracy when the LCS is assessed by our methods, in comparison with related works. As a result, the accuracies obtained in the multi-ontology scenario almost rivalled (but rarely surpassed) those obtained in an ideal mono-ontology setting, even though we considered heterogeneous sources with different scopes (WordNet and MeSH). Hence, our methods would be useful when dealing with multi-ontology similarity scenarios (concepts belonging to distinct ontologies), even though mono-ontology similarity would be preferred by its accuracy, simplicity and efficiency when both concepts appears in the same ontology.

In the future, we plan to evaluate other ontology-based similarity measures in the multi-ontology scenario by applying our methods. For measures based on more complex principles than edge-counting ones, non-trivial modifications should be introduced. For IC-based measures, the estimation of the IC of a concept across several ontologies should be normalised in a way that individual values could be compared and coherently integrated. Feature-based measures, moreover, would require more than a unique matched subsumer pair, because they rely on the whole set of common features observed for the compared terms (instead of a unique

LCS). In this case, specific graph-matching algorithms should be applied in order to obtain multiple bijections between ontologies. Moreover, due to our methods have been designed in a generic way, so that they can be applied to any structure offering a taxonomical backbone, we plan to evaluate them in other domains [52]. Finally, the ability to discover equivalent or similar concepts across different ontologies could be also exploited as a first step in ontology integration [11].

Acknowledgements

This work was partly funded by the Spanish Government through the projects CONSOLIDER INGENIO 2010 CSD2007-0004 “ARES” and CICYT DPI 2010-17112, and by the Government of Catalonia under grant 2009 SGR 1135.

References

[1] K. Saruladha, G. Aghila, A. Bhuvaneswary, Computation of Semantic Similarity among Cross Ontological Concepts for Biomedical Domain, Journal of Computing 2 (2010) 111-118.

[2] D. Bollegala, Y. Matsuo, M. Ishizuka, WebSim: A Web-based Semantic Similarity Measure, 21st Annual Conference of the Japanese Society for Artificial Intelligence, JSAI 2007, Miyazaki, Japan, 2007, pp. 757-766.

[3] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems 35 (2010) 383-413.

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

[5] N. Guarino, Formal Ontology in Information Systems, in: N. Guarino, (Ed.), 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, IOS Press, Trento, Italy, 1998, pp. 3-15.

[6] A. Valls, K. Gibert, D. Sánchez, M. Batet, Using ontologies for structuring organizational knowledge in Home Care assistance, International Journal of Medical Informatics 79 (2010) 370-387.

[7] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, 1998.

[8] M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, Journal of Biomedical Informatics 44 (2011) 118-125.

[9] T. Pedersen, S. Pakhomov, S. Patwardhan, C. Chute, Measures of semantic similarity and relatedness in the biomedical domain, Journal of Biomedical Informatics 40 (2007) 288-299.

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

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

[12]G. Pirró, A semantic similarity metric combining features and intrinsic information content, Data & Knowledge Engineering 68 (2009) 1289-1308.

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

[14]P. Resnik, Using Information Content to Evalutate Semantic Similarity in a Taxonomy, in: C.S. Mellish, (Ed.), 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448-453.

[15]J.J. Jiang, D.W. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, 1997, pp. 19-33.

[16]D. Lin, An Information-Theoretic Definition of Similarity, in: J. Shavlik, (Ed.), Fifteenth International Conference on Machine Learning, ICML 1998, Morgan Kaufmann, Madison, Wisconsin, USA, 1998, pp. 296-304.

[17]D. Sánchez, M. Batet, D. Isern, Ontology-based Information Content computation, Knowledge-based Systems 24 (2011) 297-303.

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

[19]A. Tversky, Features of Similarity, Psycological Review 84 (1977) 327-352.

[20]L. Ding, T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, J. Sachs, Swoogle: A Search and Metadata Engine for the Semantic Web, thirteenth ACM international conference on Information and knowledge management, CIKM 2004, ACM Press, Washington, D.C., USA, 2004, pp. 652-659.

[21]C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, WordNet: An electronic lexical database, MIT Press, 1998, pp. 265-283.

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

[23]J. Euzenat, P. Shvaiko, Ontology Matching, Springer Verlag, Amsterdam, 2007.

[24]Y. Bishr, Semantic Aspects of Interoperable GIS, Ph.D. Thesis, in: W.A.U.a. ITC, (Ed.), 1997.

[25]P. Jaccard, Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines, Bulletin de la Société Vaudoise de Sciences Naturelles 34 (1901) 241-272.

[26]H. Al-Mubaid, H.A. Nguyen, A cluster-based approach for semantic similarity in the biomedical domain, 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006 IEEE Computer Society, New York, USA, 2006, pp. 2713–2717.

[27]A. Gómez-Pérez, M. Fernández-López, O. Corcho, Ontological Engineering, 2nd ed., Springer-Verlag, 2004.

[28]G. Pirrò, M. Ruffolo, D. Talia, SECCO: On Building Semantic Links in Peer to Peer Networks, in: S. Spaccapietra, (Ed.), Journal on Data Semantics XII, Springer Berlin / Heidelberg, 2009, pp. 1-36.

[29]A. Blank, 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, Mouton de Gruyter, Berlin, Germany, 2003, pp. 37-66.

[30]A. Ochiai, Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions, Bulletin of the Japanese Society for Fish Science 22 (1957) 526-530.

[31]G.G. Simpson, Notes on the measurement of faunal resemblance, American Journal of Science 258-A (1960) 300-311.

[32]L.R. Dice, Meaures of the amount of ecologic association between species, Ecology 26 (1945) 297-302.

[33]J. Braun-Blanquet, Plant sociology: The study of plants communities, Oxford University Press, London, 1932.

[34]K. Riesen, H. Bunke, Approximate graph edit distance computation by means of bipartite graph matching, Image and Vision Computing 27 (2009) 950–959.

[35]D. Emms, R.C. Wilson, E.R. Hancock, Graph matching using the interference of discrete-time quantum walks, Image and Vision Computing 27 (2009) 934-949.

[36]J. Konc, D. Janežič, A Branch and Bound Algorithm for Matching Protein Structures, 8th international conference on Adaptive and Natural Computing Algorithms, 2007, pp. 399-406.

[37]S. Melnik, H. Garcia-Molina, E. Rahm, Similarity Flooding: A Versatile Graph Matching Algorithm, 18th International Conference on Data Engineering, 2002, pp. 117-128.

[38]P. Doshi, R. Kolli, C. Thomas, Inexact Matching of Ontology Graphs Using Expectation-Maximization, Journal of Web Semantics 7 (2009) 90-106.

[39]R. Myers, R.C. Wilson, E.R. Hancock, Bayesian Graph Edit Distance, Pattern Analysis and Machine Intelligence 22 (2000) 628-635.

[40]G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Language and Cognitive Processes 6 (1991) 1-28.

[41]H. Rubenstein, J. Goodenough, Contextual correlates of synonymy, Communications of the ACM 8 (1965) 627-633.

[42]A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis, E.E. Milios, Information Retrieval by Semantic Similarity, International Journal on Semantic Web and Information Systems 2 (2006) 55-73.

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

[44]S.J. Nelson, D. Johnston, B.L. Humphreys, Relationships in Medical Subject Headings, Relationships in the Organization of Knowledge, K.A. Publishers, 2001, pp. 171-184.

[45]S. Patwardhan, S. Banerjee, T. Pedersen, Using Measures of Semantic Relatedness for Word Sense Disambiguation, in: A.F. Gelbukh, (Ed.), 4th International Conference on Computational Linguistics and Intelligent Text Processing and Computational Linguistics, CICLing 2003, Springer Berlin / Heidelberg, Mexico City, Mexico, 2003, pp. 241-257.

[46]A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA, 2001, pp. 10-15.

[47]R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering 19 (2006) 370-383.

[48]M. Stevenson, M.A. Greenwood, A semantic approach to IE pattern induction, in: K. Knight, (Ed.), 43rd Annual Meeting on Association for Computational Linguistics, COLING-ACL 2005, Association for Computational Linguistics, Ann Arbor, Michigan, USA, 2005, pp. 379–386.

[49]D. Sánchez, D. Isern, Automatic extraction of acronym definitions from the Web, Applied Intelligence 34 (2011) 311-327.

[50]D. Sánchez, D. Isern, M. Millán, Content Annotation for the Semantic Web: an Automatic Web-based Approach, Knowledge and Information Systems 27 (2010) 393-418.

[51]D. Sánchez, A. Moreno, Learning non-taxonomic relationships from web documents for domain ontology construction, Data & Knowledge Engineering 63 (2008) 600-623.

[52]D. Sánchez, A methodology to learn ontological attributes from the Web, Data & Knowledge Engineering 69 (2010) 573-597.

[53]D. Sánchez, A. Moreno, Pattern-based automatic taxonomy learning from the Web, AI Communications 21 (2008) 27-48.

[54]J.E. Caviedes, J.J. Cimino, Towards the development of a conceptual distance metric for the UMLS, Journal of Biomedical Informatics 37 (2004) 77-85.

[55]H.-M. Lu, H. Chen, D. Zeng, C.-C. King, F.-Y. Shih, T.-S. Wu, J.-Y. Hsiao, Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints, International Journal of Medical Informatics 78 (2009) 308-320.

[56]G. Papachristoudis, S. Diplaris, P.A. Mitkas, SoFoCles: Feature filtering for microarray classification based on Gene Ontology, Journal of Biomedical Informatics 43 (2010) 1-14.

[57]G. Nenadi, H. Mima, I. Spasi, S. Ananiadou, J.-i. Tsujii, Terminology-driven literature mining and knowledge acquisition in biomedicine, International Journal of Medical Informatics 67 (2002) 33-48.

[58]V. Sugumaran, V.C. Storey, Ontologies for conceptual modeling: their creation, use, and management, Data & Knowledge Engineering 42 (2002) 251-271.

[59]I. Bichindaritz, S. Akkineni, Concept mining for indexing medical literature, Engineering Applications of Artificial Intelligence 19 (2006) 411-417.

Appendix A

This appendix includes individual similarity results for the two evaluation benchmarks and the Rada measure (results are inverted by changing the sing to convert them to similarity values), under the mono-ontology and multi-ontology scenarios considered in the evaluation (section 6).

Table A.I. Similarity results for the different mono-ontology and multi-ontology scenarios considered in the evaluation (see section 6) for the Pedersen et al. benchmark. **Bold** columns represent our methods.

WordNet term	*MeSH term*	*Physician ratings*	*Coder ratings*	MeSH	WordNet	MeSH + WordNet Strategy1	MeSH + WordNet Strategy2	**MeSH + WordNet Strategy3**	**MeSH + WordNet Strategy4**
Renal failure	Kidney failure	4.0	4.0	0	0	-13	0	**0**	**0**
Myocardium	Heart	3.3	3.0	-1	-8	-12	-12	**-1**	**-1**
Infarct	Stroke	3.0	2.8	-8	-11	-14	-14	**-6**	**-7**
Abortion	Miscarriage	3.0	3.3	0	0	-14	-14	**0**	**0**
Schizophrenia	Delusion	3.0	2.2	-7	-4	-14	-14	**-9**	**-1**
Adenocarcinoma	Metastasis	2.7	1.8	-6	-18	-18	-4	**-4**	**-4**
Stenosis	Calcification	2.7	2.0	-7	-7	-14	-14	**-6**	**-7**
Diarrhoea	Stomach cramps	2.3	1.3	-3	-3	-14	-14	**-6**	**-6**
Atrial fibrillation	Mitral stenosis	2.3	1.3	-4	-4	-15	-4	**-4**	**-4**
Rheumatoid arthritis	Lupus	2.0	1.1	-2	-4	-15	-2	**-2**	**-2**
Osteoarthritis	Carpal tunnel syndrome	2.0	1.1	-7	-12	-17	-7	**-7**	**-7**
Hypertension	Diabetes mellitus	2.0	1.0	-5	-11	-11	-11	**-5**	**-5**
Acne	Syringe	2.0	1.0	-8	-20	-14	-14	**-9**	**-7**
Antibiotic	Allergy	1.7	1.2	-9	-17	-10	-10	**-10**	**-10**
Multiple sclerosis	Psychosis	1.0	1.0	-9	-6	-12	-12	**-4**	**-4**
Appendicitis	Osteoporosis	1.0	1.0	-8	-14	-14	-14	**-6**	**-6**
Xerostomia	Alcoholic cirrhosis	1.0	1.0	-8	-8	-12	-12	**-6**	**-6**
Peptic ulcer disease	Myopia	1.0	1.0	-6	-9	-15	-15	**-7**	**-8**
Cellulitis	Depression	1.0	1.0	-9	-11	-14	-14	**-14**	**-7**
Hyperlipidaemia	Metastasis	1.0	1.0	-8	-12	-12	-12	**-4**	**-5**

Correlations against physicians	-	-	0.66	0.39	-0.16	0.35	**0.48**	**0.59**
Correlations against coders	-	-	0.63	0.48	0	0.19	**0.55**	**0.64**

Table A.II. Similarity results for the different mono-ontology and multi-ontology scenarios considered in the evaluation (see section 6) for the Hliaoutakis et al. benchmark. **Bold** columns represent our methods.

WordNet term	*MeSH term*	*Expert ratings*	MeSH -	WordNet -	MeSH + WordNet Strategy1	MeSH + WordNet Strategy1	**MeSH + WordNet Strategy3**	**MeSH + WordNet Strategy4**
Appendicitis	Anemia	0.031	-7	-3	-13	-13	**-5**	**-5**
Otitis Media	Infantile Colic	0.156	-9	-5	-16	-16	**-8**	**-8**
Dementia	Atopic Dermatitis	0.060	-8	-12	-14	-14	**-8**	**-8**
Malaria	Bacterial Pneumonia	0.156	-6	-4	-14	-14	**-6**	**-7**
Osteoporosis	Patent Ductus Arteriosus	0.156	-8	-19	-14	-14	**-6**	**-7**
Antibacterial Agents	Amino Acid Sequence	0.155	-11	-9	-11	-11	**-11**	**-10**
Congenital Heart Defects	Acq. Immunno. Syndrome	0.060	-6	-8	-12	-12	**-7**	**-7**
Meningitis	Tricuspid Atresia	0.031	-7	-9	-17	-7	**-7**	**-7**
Sinusitis	Mental Retardation	0.031	-7	-8	-13	-13	**-6**	**-7**
Hypertension	Kidney Failure	0.500	-7	-4	-13	-13	**-7**	**-7**
Hyperlipidemia	Hyperkalemia	0.156	-5	-2	-13	-13	**-5**	**-6**
Hypothyroidism	Hyperthyroidism	0.406	-2	-2	-12	-12	**-1**	**-2**
Sarcoidosis	Tuberculosis	0.406	-10	-7	-16	-16	**-8**	**-9**
Vaccines	Immunity	0.593	-7	-11	-8	-8	**-8**	**-8**
Asthma	Pneumonia	0.375	-3	-2	-15	-5	**-2**	**-2**
Diabetes Mellitus	Diabetic Nephropathy	0.500	-2	-10	-18	-2	**-2**	**-2**
Lactose Intolerance	Irritable Bowel Syndrome	0.468	-5	-17	-18	-8	**-8**	**-8**
Urinary Tract Infection	Pyelonephritis	0.656	-5	-1	-18	-8	**-1**	**-1**
Sepsis	Neonatal Jaundice	0.187	-6	-14	-14	-14	**-6**	**-7**
Anemia	Deficiency Anemia	0.437	-2	-1	-16	-2	**-2**	**-2**
Psychology	Cognitive Science	0.593	-4	-2	-14	-1	**-1**	**-1**
Adenovirus	Rotavirus	0.437	-5	-3	-10	-4	**-4**	**-4**
Migraine	Headache	0.718	*-8*	*-1*	-16	0	**0**	**0**
Myocardial Infarction	Myocardial Ischemia	0.750	-1	-8	-14	-14	**-6**	**-7**

Hepatitis B	Hepatitis C	0.562	-2	-2	-17	-4	**-3**	**-2**
Carcinoma	Neoplasm	0.750	-3	-3	-15	-1	**-1**	**-1**
Pulmonary Stenosis	Aortic Stenosis	0.531	-2	-2	-15	-15	**-2**	**-2**
Breast Feeding	Lactation	0.843	-8	-3	-15	-15	**-3**	**-3**
Antibiotics	Antibacterial Agents	0.937	0	-1	-13	-8	**-7**	**-7**
Seizures	Convulsions	0.843	0	-1	-13	0	**0**	**0**
Ache	Pain	0.875	0	-1	-13	0	**0**	**0**
Malnutrition	Nutritional Deficiency	0.875	0	-1	-15	0	**0**	**0**
Measles	Rubeola	0.906	0	0	-19	0	**0**	**0**
Chicken Pox	Varicella	0.968	0	0	-18	0	**0**	**0**
Down Syndrome	Trisomy 21	0.875	0	0	-13	0	**0**	**0**
Correlations against human experts		-	0.68	0.53	-0.21	0.63	**0.65**	**0.67**