Semantic similarity estimation from multiple ontologies

Montserrat Batet[1]*, David Sánchez[1], Aida Valls[1], Karina Gibert[2]

(1) *Universitat Rovira i Virgili. Departament d'Enginyeria Informàtica i Matemàtiques.Av. Països Catalans, 26 (Campus Sescelades). 43007 Tarragona, Catalonia (Spain)*

(2) *Universitat Politècnica de Catalunya. Department of Statistics and Operational Research.C/ Jordi Girona, 1-3. 08034 Barcelona, Catalonia (Spain)*

Abstract

The estimation of semantic similarity between words is an important task in many language related applications. In the past, several approaches to assess similarity by evaluating the knowledge modelled in an ontology have been proposed. However, in many domains, knowledge is dispersed through several partial and/or overlapping ontologies. Because most previous works on semantic similarity only support a unique input ontology, we propose a method to enable similarity estimation across multiple ontologies. Our method identifies different cases according to which ontology/ies input terms belong. We propose several heuristics to deal with each case, aiming to solve missing values, when partial knowledge is available, and to capture the strongest semantic evidence that results in the most accurate similarity assessment, when dealing with overlapping knowledge. We evaluate and compare our method using several general purpose and biomedical benchmarks of word pairs whose similarity has been assessed by human experts, and several general purpose (WordNet) and biomedical ontologies (SNOMED CT and MeSH). Results show that our method is able to improve the accuracy of similarity estimation in comparison to single ontology approaches and against state of the art related works in multi-ontology similarity assessment.

Keywords

Semantic similarity, Ontologies, knowledge representation, WordNet, MeSH, SNOMED.

* Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain
Tel.: +34 977 556563; Fax: +34 977 559710;
E-mail: montserrat.batet@urv.net.

1 Introduction

With the enormous success of the Information Society, the amount of textual electronic information available has been significantly increasing in recent years. As a result, computer understanding of electronic texts has become an important trend in computational linguistics. One of the most basic tasks is the evaluation of the semantic similarity between words. Semantic similarity/distance methods have been extensively developed to tackle this problem in an automatic way. Word similarity estimation has many direct applications. In word-sense disambiguation [1], for example, context terms can be semantically compared with the senses of a potentially ambiguous word to discover the most *similar* sense. In document categorisation or clustering [2], the semantic resemblance between words can be compared to group documents according to their subject. In word spelling correction [3], semantic similarity can assess which is the most appropriate correction for a potential misspelling according to its similarity against context words (e.g., "a *cot* is a mammal" instead of "a *cat* is a mammal"). Automatic language translation [2] relies on considering the detection of terms pairs expressed in different languages but referring to the same concept as a synonym discovery task, where semantic similarity assists the detection of different linguistic formulations of the same concept. Semantic similarity assessments can also assist information extraction tasks [4] such as semantic annotation [5] and ontology learning [6-10], helping to discover semantically related terms. Finally, semantic similarity is widely used in information retrieval [3,11-13,4] tasks, either suggesting similar queries to improve the recall or filtering results according to their resemblance to the user query.

Semantic similarity is estimated from the degree of taxonomic proximity between concepts. For example, *bronchitis* and *flu* are similar because both are respiratory system disorders. Taxonomies, and more general ontologies, provide a formal and machine-readable way to express a shared conceptualisation by means of a unified terminology and semantic inter-relations [14], from which similarity can be estimated. Motivated by initiatives such as the Semantic Web [15], many ontologies have been developed in the last years. Available ontologies range from general-purpose knowledge sources, such as WordNet [16] for English words, to specific terminologies, such as medical sources like UMLS, or ontologies designed for a specific application [17-20]. Thanks to the explicit knowledge structure that ontologies provide, they have been extensively used to compute similarity. In fact, similarity estimation is based on the extraction of semantic evidence from one or several knowledge sources. The more available the background knowledge is and the better its structure is, the more accurate the estimation will potentially be. Different families of measures can be disguised according to the type of knowledge used to extract semantic evidences and to the principles in which similarity estimation relies.

On the one hand, there exist pure ontology-based methods focused on the analysis of the hierarchical structure of an ontology [21-24]. Ontologies are considered in these approaches as geometrical models in which inter-concept similarity can be estimated from their relative *distance* in the ontological structure, which is the shortest *path* of semantic links connecting them [24]. On the other hand, other similarity computation paradigms grounded in the information theory complement the taxonomic knowledge provided by an ontology with the probability of appearance of words in a pre-processed domain corpus [25-27]. However, these latter approaches depend on corpora availability and human tagging to obtain robust similarity estimations [28-30]. On the contrary, pure ontology-based approaches are able to compute similarity in an efficient manner without depending on external resources and human supervision [31]. In fact, they are able to provide accurate results when a well detailed and taxonomically homogenous ontology is available [23]. This is also a drawback, because they completely depend on the degree of coverage and detail of the input ontology. This limitation could be overcome by using multiple ontologies.

Classical similarity approaches, in general, do not support more than one input ontology. The use of multiple ontologies provide additional knowledge [32] that may help to improve the similarity estimation and to solve cases in which terms are not represented in a certain ontology. This is especially interesting in domains in which several large and detailed ontologies are available (e.g., in biomedicine: MeSH and SNOMED) offering overlapping and/or complementary knowledge of the same domain. However, the different scopes, points of view and design principles followed by knowledge experts when developing ontologies produce significant differences in their level of detail, granularity and semantic structure, making difficult the comparison and integration of similarities computed from different ontologies [32].

As it will be detailed in the section 3, very little work has been done in developing methods to enable the similarity assessment from multiple ontologies. In this paper we present a new approach tackling this task. On the one hand, our method will permit estimating the similarity when a term is missing in a certain ontology but it is found in another one by discovering common concepts that can act as bridges between different knowledge sources. On the other hand, in case of overlapping knowledge, that is, ontologies covering the same pair of terms, our approach will be able to improve the accuracy by selecting the, apparently, most reliable similarity estimation from those computed from the different ontologies. A heuristic has been designed to tackle this task, based on the amount of semantic evidences observed for each ontology. It is worth to note that, on the contrary to previous works, our approach tackles the cases found in a multi-ontology scenario according to which ontology the concepts belong (i.e., the pair of

compared concepts belong to one ontology, both concepts are found in more than one ontology, each concept belongs to a different ontology) and it operates in an unsupervised fashion during the semantic integration. Our method has been applied to a state-of-the-art ontology-based measure [31], designed to provide a high similarity estimation accuracy by exploiting solely taxonomical knowledge.

The rest of the paper is organised as follows. Section 2, provides some definition of basic concepts involved in the semantic similarity assessment. Section 3 introduces related works focusing on ontology-based semantic similarity using multiple ontologies. Section 4 begins with a description of the similarity measure to which the proposed method has been applied. Then, the method for similarity assessment from several ontologies is described in detail, defining a set of cases according to which ontologies the input terms belong and the strategy proposed to tackle each situation. Section 5 evaluates our approach by means of several standard benchmarks and compares it against a mono-ontology scenario and related works. The final section contains the conclusions.

2 Basic Concepts

Let us define $path(c_1, c_2)=\{l_1,....l_k\}$ as the minimum number of links connecting the terms c_1 and c_2 in a taxonomy. Let $|path(c_1, c_2)|$ the length of this path.

Example 1. Let us consider the ontology O_1 which is partially shown in Figure 1.

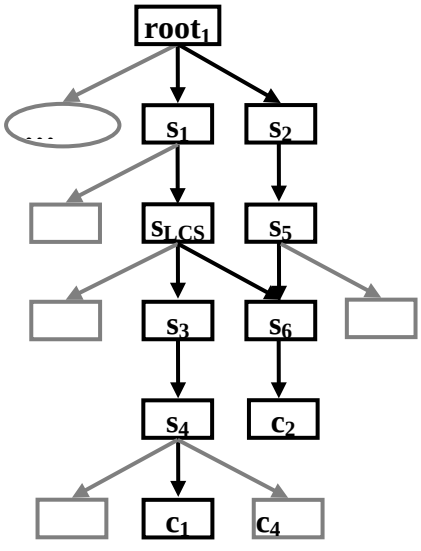

Figure 1: Ontology O_1

Given a pair of concepts c_1, c_2 in O_1, we can see that the ~~shortest~~ path between c_1, and c_2 is composed by the set of links that connect them: $path(c_1, c_2)=\{c_1\text{-}s_4, s_4\text{-}s_3, s_3\text{-}s_{LCS}, s_{LCS}\text{-}s_6, s_6\text{-}c_2\}$. Consequently, $|path(c_1, c_2)|=5$.

Applying *path* definition, given a concept *c*, we can define *depth*(*c*) as the length of the path between *c* and the *root* node (1).

$$depth(c) = | path(c, root) | \quad (1)$$

Considering Example 1, *depth* (c_1) = 5.

In addition, let us define a superconcept *s* of a concept *c* as any concept that taxonomically subsumes (generalises) *c* in the taxonomy.

Applied to Example 1, concepts s_4, s_3, s_{LCS}, s_1 and the *root* node are superconcepts of c_1.

Finally, let us consider *Sup*(c_1, c_2) the set of superconcepts that subsume both c_1 and c_2. Then, the *Least Common Subsummer* (*LCS*) of concepts c_1 and c_2 (i.e. *LCS*(c_1, c_2)) is defined as the most specific superconcept (the one with the maximum depth) that defines the path between the pair of evaluated concepts.

$$LCS(c_1, c_2) = \underset{s \in \left\{ \underset{c \in Sup(c_1, c_2)}{\arg\min} \left(|path(c_1, c)| + |path(c, c_2)| \right) \right\}}{\arg\max} Depth(s) \quad (2)$$

Applied to Example 1, the LCS of c_1 and c_2 is s_{LCS}.

3 Related work

As stated in the introduction, the exploitation of multiple input sources would lead to a better coverage and more robust similarity estimations. In the past, the general approach to data integration has been mapping the local terms of distinct ontologies into an existent single one [33-35,14] or creating a new ontology by integrating existing ones [36,37,34]. However, manual or semi-automatic ontology integration represents a challenging problem, both from the cost and scalability points of view (e.g. requiring the supervision of an expert) and because of the difficulty to deal with overlapping concepts and inconsistencies across ontologies [38].

Tackling the problem from a different perspective, [38] compute the similarity between terms as a function of some ontological features and the degree of generalisation between concepts (i.e., the path distance between the pair of concepts and their least common subsumer) into the same or different ontologies. Similarly, in [39], authors rely on the matching between synonym sets and concept definitions, using the Jaccard index as to measure the degree of overlapping and, hence, of similarity. When the term pair belongs to different ontologies, both methods approximate the degree of generalisation between concepts by considering that these ontologies are connected by a new imaginary root

node that subsumes the root nodes of these two ontologies. A problem of their approaches is their reliance on many ontological features such as attributes, synonyms, meronyms and other kind of non-taxonomic relationships, which are more scarcely found in ontologies, in comparison to taxonomical knowledge. In fact, an investigation of the structure of existing ontologies via the Swoogle ontology search engine [40] has shown that domain ontologies very occasionally model non-taxonomic knowledge. Moreover, these methods do not consider the case in which the term pair is found in several ontologies, which is a very common situation as it will be shown during the evaluation. In consequence, they omit the problem of selecting the most appropriate assessment and/or the integration of overlapping sources of information. In addition, the integration of different ontologies is very simple and does not consider the case in which ontologies share subsumers that could be used as bridging classes.

Other methods rely on terminological matchings to enable multi-ontology similarity assessment. In [41], authors compute the similarity as function of the concreteness of the concept subsuming the compared terms in a taxonomy. In the multi-ontology scenario, this subsuming concept is obtained by matching concept labels of different ontologies and retrieving the most taxonomical specific matching. A more elaborated approach is presented in [42]. This work complements terminological matching of subsumer concepts with a graph-based ontology alignment method that aims at discovering structurally similar, but not necessarily terminologically identical, subsumers. The result of the alignment method is evaluated by means of path-based similarity measures. The main problem of these methods is their omission of the case in which overlapping knowledge is available.

A more general proposal covering all possible situations which may appear in a multi-ontology scenario is presented by [32]. They apply it to the UMLS biomedical source, in which concepts are dispersed across several overlapping ontologies and terminologies such as MeSH or SNOMED (see section 5). Authors propose a methodology to exploit these knowledge sources using a path-based distance defined in [43]. The proposed measure combines *path length* and *common specificity*. They use the common specificity to consider that pairs of concepts at a lower level of a taxonomy should be more similar than those located at a higher level. They measure the common specificity of two concepts by subtracting the depth of their Least Common Subsumer from the depth D_c of the taxonomic branch to which they belong.

$$CSpec(c_1, c_2) = D_c - depth(LCS(c_1, c_2)) \quad (3)$$

The smaller the common specificity of two concept nodes, the more the information they share, and thus, the more similar they are. Based on path length and common specificity, they proposed a measure (*SemDist*) defined as follows:

$$SemDist(c_1, c_2) = \log((|\ path(c_{1,} c_2)\ | -1)^{\alpha} \times (CSpec(c_1, c_2))^{\beta} + k)\ , \qquad (4)$$

where $\alpha>0$ and $\beta>0$ are contribution factors of two features, k is a constant, and $|path(c_1,c_2)|$ is the length of the ~~shortest~~ path between the two concepts. To ensure the function is positive and the combination is non-linear, k must be greater or equal to one.

For the case of multiple ontologies, they rely on a user-selected *primary* ontology (the other ones are considered as *secondary*) that acts as the master in cases in which concepts belong to several ontologies. It is also used as a base to normalise similarity values. They propose different strategies according to the situation in which the compared concepts appear. If both concepts appear in the *primary* ontology, the similarity is computed exclusively from that source even in the case that they also appear in a *secondary* ontology. If both concepts appear only in a unique *secondary* ontology, obviously, the similarity is computed from that source. A more interesting case occurs when concepts appear in several *secondary* ontologies. Authors propose a heuristic to choose from which of these ontologies the similarity should be computed, based on the degree of overlapping with respect to the *primary* ontology and the degree of detail of the taxonomy (granularity). Finally, if a concept is uniquely found in an ontology (the *primary*) and the other concept in a different ontology (a *secondary* one), they temporally "connect" both ontologies by finding "common nodes" (i.e. subsumers representing the same concepts in any of the ontologies) and considering the result as a unique ontology.

A problem that authors face is the fact that different ontologies may have different granularity degrees, that is, different depths and branching factors for a certain taxonomic tree. Because their measure is based on absolute path lengths, the similarity computed for each term pair from different ontologies will lead to a different similarity scale which cannot be directly compared. They propose a method to scale similarity values both in the case in which the concept pair belongs to a unique *secondary* ontology or when it belongs to different ontologies - both *secondary*, or one *primary* and the other *secondary* - which are "connected", taking as reference the predefined *primary* ontology. They scale both *Path* and *CSpec* features to the *primary* ontology according to difference in depth with respect to the *primary* ontology. For example, in the case in which both concepts belong to a unique *secondary* ontology, *Path* and *CSpec* are computed as stated in (5) and (6) respectively, and they compute the similarity using (4).

$$|\ Path(c_1, c_2)\ | = |\ Path(c_1, c_2)_{secondary\_onto}\ | \times \frac{2D_1 - 1}{2D_2 - 1} \qquad (5)$$

$$CSpec(c_1, c_2) = CSpec(c_1, c_2)_{secondary\_onto} \times \frac{D_1 - 1}{D_2 - 1}\ , \qquad (6)$$

where D_1 and D_2 are the depths of the *primary* and *secondary* ontologies respectively. Hence, (D_1 -1) and (D_2 -1) are the maximum common specificity values of the primary and secondary ontologies respectively, and ($2D_1$ -1) and ($2D_2$ -1) are the maximum path values of two concept nodes in the *primary* and *secondary* ontologies, respectively.

This approach has some drawbacks. First, since the proposed method bases the similarity assessment on the *path length* connecting concept pairs, it omits other taxonomic knowledge already available in the ontology such as the complete set of common and non-common superconcepts. Moreover, it is worth to note that the presence of an *is-a* link between two concepts gives an evidence of a relationship but not about the degree of their semantic similarity, because all individual links have the same length and, in consequence, represent uniform distances [44]. Secondly, the cross-ontology method is hampered by the fact that a *primary* ontology should be defined a priori by the user to solve cases with overlapping knowledge and to normalise similarity values. This scaling process, motivated by the fact of basing the similarity on absolute values of semantic features (path and depth) of a concrete ontology, results in a complex scenario to be considered during the similarity assessment. Moreover, it assumes that, in all situations, the *primary* ontology will lead to better similarity estimations than *secondary* ones, which could not be true. There may be situations in which, for a pair of concepts appearing both in the *primary* ontology and also in one or several *secondary* ontologies, similarity estimation from a *secondary* one may lead to better results because its knowledge representation is more accurate or more detailed for a particular subset of terms. Even though authors evaluate the approach using standard benchmarks and widely available ontologies, experiments regarding the influence in the results of selecting one ontology or another as the *primary* are missing.

4 Proposed method

In section 4.1, we provide a brief description of a similarity measure selected as more appropriate way to estimate similarity than related works based on absolute path-lengths [32,38]. In section 4.2, a multi-ontology similarity enabling method is presented jointly with the heuristics proposed to tackle the different scenarios, that is, when overlapping or complementary knowledge is available, and on the contrary to most related works [39,38,41,45] focusing only on the former case.

4.1 Similarity measure

From section 3 we realise that one of the problems of the approaches presented by [32] and [38] is the similarity function used in their methodologies. Regarding [32], being a

path-based function, their underlying measure provides absolute similarity values with non-comparable scales when they are obtained from different ontologies, that is, the path length would depend on the ontology size, depth and granularity. In the case of the measure used by [38] their measure relies on non-taxonomic features, which are rarely found in ontologies [40].

From these considerations, we conclude that an ontology-based measure providing relative values normalised according to the ontological structure to which it has been applied, that it is based solely on taxonomical knowledge is desirable in a multi-ontology scenario.

Several ontology-based measures fitting these requirements exist [21,23,31]. Those were compared in [31], concluding that the approach by [31] was able to provide the best results when evaluated with standard ontologies and benchmarks. Particularly, this measure tackles some of the limitations observed in other ontology-based measures relying on the quantification of the path [21,23]. Because of their simplicity, these approaches consider partial knowledge (i.e., the path and/or taxonomic depth) during the similarity assessment. This omits other taxonomic relations defined between concepts in case, for example, of ontologies incorporating relations of multiple inheritance in which several paths between concept pairs exist. The approach by [31], on the contrary, evaluates concept similarity as a function of the amount of common and non-common taxonomic subsumers of the compared concepts. Concretely, it evaluates, in a non-linear way, the ratio between the cardinality of the set of non-common superconcepts as an indication of distance, and the total number of superconcepts of both concepts as a normalising factor (7). In this manner, it exploits more ontological knowledge than path-based measures since, if multiple taxonomic superconcepts exist, all of them are considered. This improves the accuracy in a mono-ontology setting, but retaining the computational simplicity and lack of constraints [31]. The measure was defined as:

$$sim(c_1, c_2) = -\log_2 \frac{\left|T(c_1) \cup T(c_2)\right| - \left|T(c_1) \cap T(c_2)\right|}{\left|T(c_1) \cup T(c_2)\right|}, \qquad (7)$$

where $T(c_i)$ is defined as the set of superconcepts of the concept c_i, including the concept c_i, as $T(c_i)=\{c_j \in C \mid c_j$ is superconcept of $c_i\} \cup \{c_i\}$.

As shown in (7), the fact that it evaluates a ratio of taxonomic features provides normalised similarity values which can be compared independently of the ontology size and granularity. This is relevant in a multi-ontology setting because it will enable a direct comparison of the results obtained from different ontologies.

Because of these arguments, we selected this measure as the function to which apply and test the accuracy of the multi-ontology similarity computation method presented below.

4.2 Multi-ontology similarity assessment

Our multi-ontology similarity computation method has been designed to be general enough to be applicable to any ontology configuration without requiring user supervision. On the contrary to [38,39,45,41], our method considers all possible situations according to which ontology the compared concepts belong. On the contrary to [32], who rely on the pre-selection of a *primary* ontology, we consider all input ontologies *equally important.* As a result, the multi-ontology similarity scenario is simplified to three cases (instead of five proposed in [32]), according to whether or not each or both concepts (c_1 and c_2) belong to any of the considered ontologies.

Case 1: *Concepts c_1 and c_2 appear in only one ontology*.

If the pair of concepts occurs in a unique ontology, the similarity is computed like in a mono-ontology setting.

Example 2. Let us compare the concept pair c_1 and c_2 belonging to a unique ontology O_1, whose structure is partially shown in Figure 1. Applying the measure detailed in (7), so that we obtain: $T(c_1) \cap T(c_2)$={root, s_1, s_{LCS}} and $T(c_1) \cup T(c_2)$= {$root_1$, s_1, s_{LCS}, s_2, s_3, s_4, s_5, s_6, c_1, c_2}, then sim(c_1,c_2)= -$\log_2$((10-3)/10)=0.514.

Case 2: *Both concepts c_1 and c_2 appear at the same time in more than one ontology*.

In this case, both concepts appear in several ontologies, each one modelling knowledge in a different but overlapping way. Hence, the similarity calculus is influenced by the different levels of detail or knowledge representation accuracy of each ontology [38].
A possible way to tackle this situation would be to aggregate this overlapping knowledge or the individual similarity estimations pursuing the hypothesis that the combination of individual evidences would produce more accurate results. However, when dealing with ontological models, as acknowledged by researchers working on ontology merging/alignment [46], the integration of heterogeneous knowledge becomes a complex task due individual ontologies are typically created by different experts, pursuing different goals, framed in different areas/tasks and with different scopes and points of view. Hence, even in cases in which overlapping knowledge is detected across several ontologies like in this case, the aggregation of this knowledge is challenging because of ambiguities and semantic inconsistencies. As a result, the integration of different knowledge sources cannot rival the semantic coherency of a knowledge structure created by an individual expert [46].

In this case, we are dealing with several knowledge structures that model the same knowledge (concept pairs), each one being created by an individual expert in a semantically coherent manner. Hence, similarity assessments from individual ontologies would be also coherent. Considering the difficulties and imperfections of automatic knowledge integration processes, we argue that individual similarity assessments computed from individual ontologies would be more accurate.

Moreover, when dealing with multiple ontologies, a particular ontology may better represent similarity between concepts with respect to the rest of ontologies due to a higher taxonomical detail or a better knowledge representation accuracy.

Hence, it would be desirable to be able to decide, according to a heuristic, which ontology provides the, apparently, best estimation of the inter-concept similarity. Considering the nature of the ontology engineering process, and the psychological implications of a human assessment of the similarity, two premises can be enounced.

First, the fact that a pair of concepts obtains a high similarity score is the result of considering common knowledge modelled through an explicit ontology engineering process. However, because of the knowledge modelling bottleneck, which typically affects manual approaches, ontological knowledge is usually partial and incomplete [47]. As a result, if two concepts appear to be semantically far because they, apparently, do not share knowledge, one cannot ensure if this is an implicit indication of semantic disjunction or the result of partial or incomplete knowledge. We can conclude that explicitly modelled knowledge is more important as semantic evidence when computing similarity than the lack of it. Hence, we prefer higher similarity values because they are based on explicit evidences provided by the knowledge expert.

Secondly, psychological studies have demonstrated that humans pay more attention to similar than to different features during the similarity assessment [33,48]. Hence, we assume that non-common characteristics between entities are less important than common ones.

As a result, given a pair of concepts appearing in different ontologies, we consider the one giving the *highest* similarity score as the most reliable estimation because it incorporates the highest amount of explicit evidences of relationship between terms (explicitly modelled by the knowledge engineer). Concretely, we compute similarity values individually for each ontology. Then, we choose the highest similarity value, which we assume to be the best estimation, as the final result (8).

$$sim(c_1, c_2) = \max_{\forall O_i \in O | c_1, c_2 \in O_i} sim_{O_i}(c_1, c_2), \qquad (8)$$

being O a set of ontologies to which c_1 and c_2 belong.

Applying this heuristic, given a set of concept pairs belonging to several ontologies, the final similarity value of each pair may be taken from different ontologies. This will correspond to cases in which a particular ontology provides, apparently, a more accurate modelling of the two concepts, regardless its global degree of granularity or detail in comparison to the other ontologies. On the contrary to the method by [32] that relies on the user criteria to pre-select the most adequate ontology (i.e. the *primary* one) assuming that it will provide the most accurate assessments in all situations, which is hardly assumable, our heuristic will exploit the benefits offered by each ontology for each pair of concepts. This configures an unsupervised and adaptive method that is able to select, for each term pair to compare, the assessment which is likely to more accurately represent their similarity.

Moreover, because the underlying similarity measure (7) provides *relative* values, normalised to the granularity degree of each ontology, the comparison between the results obtained from different ontologies for the same pair of concepts does not require additional scaling, on the contrary to [32]. This simplifies the process, avoiding the necessity that the user provides the normalising factor (i.e. the *primary* ontology).

Example 3. Let $O=\{O_1,O_2\}$ (see Figure 2) the set of ontologies to which c_1 and c_2 belong. Hence, in O_1, $\text{sim}_{O1}(c_1,c_2)= -\log_2((10-3)/10)=0.514$, while in O_2, $sim_{O2}(c_1,c_2)= -\log_2((9-3)/9)=0.585$. Following the heuristic (8), the final similarity is $sim(c_1,c_2)= 0.585$.

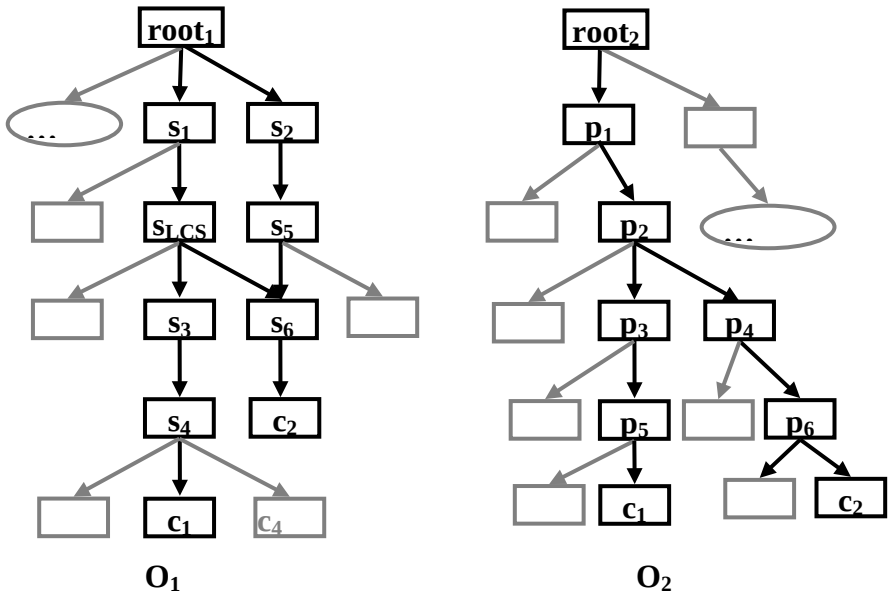

Figure 2: Ontologies O_1 and O_2

Case 3: None of the ontologies contains concepts c_1 and c_2 simultaneously

Each of the two concepts belong to a different ontology, each one modelling the knowledge from a different point of view. As stated in [38] the similarity estimation across ontologies can be only achieved if they share some components. On the contrary to the previous case, the current scenario necessarily requires integrating different ontologies to measure the similarity across them.

As stated in section 3, some approaches tackled this problem by merging different ontologies in a unique one, introducing a high computational and human cost, and dealing with the difficulties inherent to the treatment of ambiguous overlapping concepts and to the avoidance of inconsistencies [38,39,41,45].

From a different point of view, as introduced in section 3, [32] base their proposal in the differentiation between *primary* and *secondary* ontologies, connecting the *secondary* to the *primary* by joining all the equivalent nodes, that is, those with concepts represented by the same *label*. These equivalent nodes are called *bridges*. Then, they define the LCS of a pair of concepts in two ontologies as the LCS of the concept belonging to the *primary* ontology and one of the *bridge* nodes.

$$LCS_n(c_1, c_2) = LCS(c_1, bridge_n) \quad (9)$$

Then, the path is computed trough the two ontologies via the LCS and the *bridge* node, and the similarity is assessed. Again, because ontologies have different granularity degrees, the path length depends on the concrete ontology. To normalise the value, it is measured the path between the concept of the *primary* ontology and the LCS, and the path between the concept of the *secondary* ontology and the LCS *scaled* with respect to the dimension of the *primary* ontology.

In [38,39], the two ontologies are simply connected by creating a new node (called *anything*) which is a direct superconcept of their roots.

Considering the nature of our similarity measure, in which the set of common and non-common superconcepts are evaluated, the approach proposed in [38,39] implies the loss of all the potentially common superconcepts. Moreover, differently from [32,45], where only the path length to the LCS is computed, we need a more complete vision of the taxonomic structure above the evaluated concepts, including all taxonomic relationships in cases of multiple inheritance.

Because of these reasons, we propose a new method to assess similarity across different ontologies. It is based on evaluating the union of the set of superconcepts of c_1 and c_2 in each ontology and on finding equivalences between them. This allows evaluating the amount of common and non-common knowledge in a cross-ontology setting. It is important to note that, on the contrary to [32], where all the bridge nodes of the ontology are considered introducing a high computational burden in big ontologies, we only evaluate the superconcepts of each concept.

The detection of equivalencies between concepts of different ontologies has been previously studied in the *ontology alignment* field [49]. Several approaches have been proposed, based on different principles, to assess the chance that concepts of different ontologies are in fact equivalent. Many of these methods rely on semantic similarity functions to enable this assessment, provided by human experts or computed from other knowledge sources. However, in a scenario such as the current one, in which neither user intervention nor additional knowledge other than the one to be aligned is available, an unsupervised method is needed. *Terminological* matching methods, which are also widely used in the ontology alignment field [50,51,46] fit with these requirements, because they discover equivalent concepts relying solely on the fact that concept *labels* match. Applied to our problem we discover equivalent superconcepts when they are referred with the same textual labels, considering, if available, synonym sets.

However, because of language ambiguity (synonymy and polysemy) and differences in the knowledge representation process, a terminological matching offers a limited recall. To minimise this problem, in addition to consider common superconcepts as those that terminologically match, we also consider that all their subsumers are also common, regardless having or not an identical label. In fact, each of the evaluated concepts inheriting from terminologically equivalent superconcepts recursively inherits from all superconcepts' subsumers.

Summarising, the set of shared superconcepts for c_1 belonging to ontology O_1 and c_2 belonging to the ontology O_2 , is composed by those superconcepts of c_1 and c_2 with the same label, and also the subsumers of these equivalent superconcepts.

Formally, $T_{O1}(c_1)$ is the set of superconcepts of concept c_1 (including c_1) in the hierarchy H^C_{O1} of concepts (C_{O1}) in ontology O_1, and $T_{O2}(c_2)$ is the set of superconcepts of concept c_2 (including c_2) in the hierarchy H^C_{O2} of concepts (C_{O2}) in ontology O_2.

Let $>$ a relation C×C called generalisation ($c_i > c_j$, means that c_i is a generalisation of c_j or, in other words c_i is a superconcept of c_j). Then, we define the set of superconcepts of c_1, c_2 as:

$$T_{O1}(c_1) = \{c_i \in C_{O1} \mid c_i > c_1\} \cup \{c_1\} \quad (10)$$

$$T_{O2}(c_2) = \{c_j \in C_{O2} \mid c_j > c_2\} \cup \{c_2\} \quad (11)$$

Then, the set of terminologically equivalent superconcepts (*ES*) in $T_{O1}(c_1) \cup T_{O2}(c_2)$ is defined as:

$$ES = \{c_i \in T_{O_1}(c_1) \mid \exists c_j \in T_{O_2}(c_2) \wedge c_i \equiv c_j\} \quad (12)$$

In (12) "≡" means terminological match. Finally, the set of common superconcepts (*CS*) in $T_{O1}(c_1) \cup T_{O2}(c_2)$ is composed by elements in *ES* and all the superconcepts of elements in *ES*.

$$CS(c_1, c_2) = \bigcup_{\forall c_i \in ES} (T_{O_1}(c_i) \cup T_{O_2}(c_i)) \qquad (13)$$

The remaining elements in $T_{O1}(c_1) \cup T_{O2}(c_2)$ are considered as non-common superconcepts.

Once the set of common and non-common superconcepts have been defined, we are able to apply the similarity measure presented in (7).

Example 4. Let $O=\{O_1,O_3\}$ be the set of ontologies to which the compared concepts (c_4 and c_3) individually belong (Figure 3). Suppose that concepts $s_3 \in O_1$ and $r_3 \in O_3$ terminologically match ($s_3 \equiv r_3$). Only O_1 contains c_4 and only O_3 contains c_3. In this case, we obtain: ES=$\{s_3\}$ and $CS(c_4,c_3)$={ s_3, $root_1$, s_1, s_{LCS}, $root_3$,r_1, r_2}. s_3 and r_3 are considered the same concept. Then, $sim(c_3, c_4)= -\log_2((15-7)/15)=0.9$.

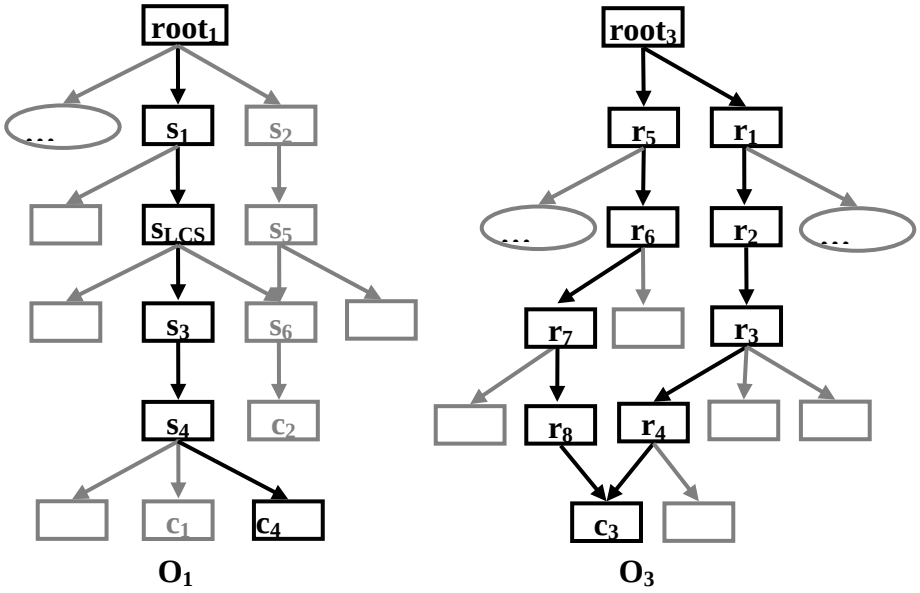

Figure 3: Ontologies O_1 and O_3

Finally, as in *Case 2*, c_1 and/or c_2 individually may belong to *several* ontologies. If that is the case, the described process and the similarity computation are executed for each combination of ontology pairs. For each pair, a similarity value is obtained. Following the same reasoning discussed in *Case 2*, we take the highest similarity values as final result. Formally:

$$sim(c_1, c_2) = \max(sim_{\{O_i, O_j\}}(c_1, c_2)), \quad \{O_i, O_j\} \in O \times O \mid i \neq j \wedge (c_1 \in O_i, c_2 \in O_j) \qquad (14)$$

Example 5. Consider the set of ontologies $O=\{O_a,O_{b,}O_{c,}O_d\}$ so that c_1 belongs to O_a and O_c and c_2 belongs to O_b and O_d. In this case, we individually compute the similarity for the pairs $O_a - O_b$, $O_a - O_d$, $O_c - O_b$ and $O_c - O_d$ using the method described above. The highest similarity value obtained for all pairs is taken as the final result.

By means of the described method approach, we are able to maintain the properties of the underlying measure, that is, the maximisation of the taxonomic knowledge exploited in the similarity assessment, but integrating knowledge from different ontologies in a seamless way. On the contrary, other approaches [32,45], which only look for a common LCS from which the path is evaluated, omit taxonomic knowledge explicitly modelled in the ontology. Again, on the contrary to [32], the numerical scale of similarity values is also maintained regardless of the input ontology because results are implicitly normalised by the size of the corresponding superconcept sets.

Although, this method could be adapted to other edge-counting measures, the similarity measure of Eq. (7) has been selected because of its desirable characteristics (discussed in section 4.1) and also because of its accuracy that surpassed other edge-counting measures and even different paradigms like corpora-based IC measure (as discussed in [31]).

5 Evaluation

Similarity measures are usually evaluated by means of standard benchmarks of word pairs whose similarity has been assessed by a group of human experts. The correlation of the similarity values obtained by computerised measures against human similarity ratings is calculated. If the correlation is near to 1, it indicates that the measure properly approximates the judgements of human subjects, which is precisely the goal.

Using widely accepted benchmarks also enables an objective comparison against other approaches. Some of these benchmarks consist of a list of domain independent pairs of words [52,53], while others have been specially designed for a specific domain. The field of biomedicine will be the focus of our evaluation in a multi-ontology scenario, as it has been very prone to the development of big and detailed ontologies and knowledge structures. The UMLS repository (Unified Medical Language System) is a paradigmatic example, which includes several biomedical ontologies and terminologies (MeSH, SNOMED CT or ICD). These ontologies are also characterised by their high level of detail, classifying concepts in several overlapping hierarchies.

In order to evaluate the accuracy of our similarity method in a multi-ontology setting and to compare it against related works, we have performed several tests combining different

biomedical and general purpose ontologies described in section 5.2. Evaluation benchmarks have been also selected accordingly; they are detailed in section 5.1.

5.1 Evaluation measures

From a domain independent point of view, the most commonly used benchmarks are those proposed by [52] and [53]. The former provides a set of 30 domain-independent word pairs manually rated by human subjects from 0 to 4. The pairs, which represent a high, middle and low level of synonymy, were chosen from an experiment done by [53], who proposed a set of manually rated 65 word pairs. Many authors [27,26,25] have used these benchmarks to evaluate and compare the accuracy of their proposals.

Within the biomedical field, we have considered two different biomedical datasets [54,55]. In the first one, [54] created, in collaboration with Mayo Clinic experts, a set of word pairs referring to general medical disorders. The similarity of each concept pair was assessed by a group of 3 physicians who were experts in the area of rheumatology and 9 medical coders who were aware about the notion of semantic similarity. After a normalisation process, a final set of 30 word pairs with the averaged similarity measures provided by both sets of experts in a scale between 1 and 4 were obtained (see Table 1). The correlation between physician judgements was 0.68, and between the medical coders was 0.78.

Table 1: Set of 30 medical term pairs with averaged experts' similarity scores (extracted from [54]).

Term 1	Term 2	Physician ratings (averaged)	Coder ratings (averaged)
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	*Lung infiltrates*	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpal tunnel syndrome	Osteoarthritis	2.0	1.1

Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulitis	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlipidemia	Metastasis	1.0	1.0

The second biomedical benchmark, proposed by [55], is composed by a set of 36 word pairs extracted from the MeSH repository (see Table 2). The similarity between word pairs was also assessed by 8 medical experts from 0 (non-similar) to 1 (synonyms).

Table 2: Set of 36 medical term pairs with averaged experts' similarity scores (extracted from [55]).

Term 1	Term 2	Human ratings (averaged)
Anemia	Appendicitis	0.031
Otitis Media	Infantile Colic	0.156
Dementia	Atopic Dermatitis	0.060
Bacterial Pneumonia	Malaria	0.156
Osteoporosis	Patent Ductus Arteriosus	0.156
Amino Acid Sequence	Antibacterial Agents	0.155
Acq. Immunno. Syndrome	Congenital Heart Defects	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
Diabetic Nephropathy	Diabetes Mellitus	0.500

Lactose Intolerance	Irritable Bowel Syndrome	0.468
Urinary Tract Infection	Pyelonephritis	0.656
Neonatal Jaundice	Sepsis	0.187
Anemia	Deficiency Anemia	0.437
Psychology	Cognitive Science	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Ischemia	Myocardial Infarction	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary Stenosis	Aortic Stenosis	0.531
Failure to Thrive	Malnutrition	0.625
Breast Feeding	Lactation	0.843
Antibiotics	Antibacterial Agents	0.937
Seizures	Convulsions	0.843
Pain	Ache	0.875
Malnutrition	Nutritional Deficiency	0.875
Measles	Rubeola	0.906
Chicken Pox	Varicella	0.968
Down Syndrome	Trisomy 21	0.875

5.2 Ontologies

We have used WordNet as domain independent ontology, and SNOMED CT and MeSH as domain-specific biomedical ontologies, because these are also used in the evaluation of previous works. Note that any other ontology (an OWL file) with a taxonomical backbone may be used instead/in addition.

WordNet [16] is a freely available lexical database that describes and structures more than 100,000 general English concepts, which are semantically structured in an ontological way. WordNet contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (synsets), each 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. To properly compare the results, we use WordNet version 2 in our tests as it is the same version used in related works.

The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT)[1] is one of the largest sources included in the Unified Medical Language System (UMLS) of the US

[1] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

National Library of Medicine. It covers most of the medical concepts, including them in one or several hierarchies. It contains more than 311,000 concepts with unique meaning organised into 18 overlapping hierarchies. SNOMED CT concepts typically present a high degree of multiple inheritance represented with approximately 1.36 million relationships.

The Medical Subject Headings (MeSH)[2] ontology is mainly a hierarchy of medical and biological terms defined by the U.S National Library of Medicine to catalogue books and other library materials, and to index articles for inclusion in health related databases including MEDLINE. It consists of a controlled vocabulary and a hierarchical tree. The controlled vocabulary contains several different types of terms such as Descriptors, Qualifiers, Publication Types, Geographic and Entry terms. MeSH descriptors are organised in a tree which defines the MeSH Concept Hierarchy. In the MeSH tree there are 16 categories, with more than 22,000 terms appearing on one or more of those categories.

5.3 Evaluation with missing terms

In the first experiment, we have taken all word pairs of the biomedical benchmarks proposed by [54] using physicians', coders' and both ratings, and [55], and the three ontologies introduced above as sources. We have evaluated each word list with several mono and multi-ontology configurations: SNOMED CT, MeSH and WordNet in an independent way (mono-ontology scenario), ontologies taken by pairs, and all of them at the same time (multi-ontology scenario).

Regarding Pedersen et al.'s benchmark, note that the term pair *"chronic obstructive pulmonary disease" - "lung infiltrates"* was excluded from the test as the latter term was not found in any of the three ontologies. For the remaining 29 pairs, all of them are contained in SNOMED CT, 25 of them are found in MeSH and 28 in WordNet. For the Hliaoutakis' benchmark, all 36 word pairs are found in MeSH and WordNet, but only 35 are contained in SNOMED CT. These values indicate that there will be some situations in which one of the words is missing in some of the ontologies but found in another. In these cases, a multi-ontology approach will potentially lead to a better accuracy, because it is able to calculate the similarity of these missing terms from the combination of multiple knowledge sources. To introduce a proper penalisation in the correlation when missing word pairs appear in a mono-ontology setting and to enable a fair comparison with regards to the multi-ontology setting, the similarity value of missing word pairs is computed in the mono-ontology setting as the average similarity for the benchmark's

[2] http://www.nlm.nih.gov/mesh/MBrowser.html

word pairs found in each ontology. Correlation values against expert's ratings obtained for all benchmarks and for all ontology combinations are shown in Table 3.

Table 3: Correlation values obtained by the proposed method for Pedersen et al.'s benchmark (with 29 word pairs) for physicians', coders' and both ratings and for Hliaoutakis' benchmark (with 36 pairs) using SNOMED CT, MeSH, and WordNet.

Ontologies	Pedersen Physicians	Pedersen Coders	Pedersen Both	Hliaoutakis
SNOMED CT	0.601	0.788	0.727	0.557
MeSH	0.562	0.769	0.694	0.749
WordNet	0.535	0.747	0.669	0.611
SNOMED CT + WordNet	0.624	0.799	0.744	0.727
MeSH + WordNet	0.596	0.790	0.724	0.770
SNOMED CT + MeSH	0.642	0.817	0.762	0.740
SNOMED CT + MeSH + WordNet	0.656	0.825	0.773	0.787

Analysing the results, we can extract several conclusions. First, we observe a surprisingly good accuracy using WordNet as ontology, especially for the Pedersen et al.'s benchmark, with correlation values that are only marginally worse than those obtained from the medical ontologies. For the Hliaoutakis' benchmark, which was designed from MeSH terms, WordNet is able to improve the correlation obtained with SNOMED CT alone. This shows that WordNet, even being a general purpose ontology, offers a good coverage of relatively common biomedical terms, possibly because parts of the WordNet taxonomy have been taken from UMLS. In fact, a 25.1% of MeSH terms are also covered in WordNet.

Secondly, we observe that, in most situations, the use of several ontologies improves accuracy in comparison to the use of ontologies individually. It is particularly interesting to see how the addition of WordNet to each medical ontology slightly improves the results. For the Pedersen et al.'s benchmark: from 0.69 to 0.72 for MeSH and from 0.72 to 0.74 for SNOMED CT. For the Hliaoutakis' benchmark: from 0.75 to 0.77 for MeSH and from 0.56 to 0.73 for SNOMED CT. This means that, at least, parts of WordNet taxonomy better correlate with human judgements than, potentially overs-specified, hierarchies of SNOMED CT or MeSH. The relative improvement obtained from the combination of the two medical ontologies (SNOMED CT and MeSH) also leads to a higher accuracy in most situations. This is reasonable because of the biomedical nature of evaluated word pairs. Finally, the combination of all ontologies provides the highest correlation in all cases (the correlation obtained against the Pedersen et al.'s medical coders is 0.825 when using all the ontologies vs. 0.788 when using only SNOMED CT, 0.769 when using MeSH and 0.747 when using WordNet). These results show that the

more available knowledge, the better the estimations there will be. This is motivated both for the resolving of missing values and thanks to the selection of the most accurate assessment from those provided by each overlapping ontology.

We can also observe that our method and, in consequence, the underlying similarity measure, correlates better with coders than with physicians for the Pedersen et al.'s benchmark. On the one hand, this is motivated by the higher amount of discrepancies observed in physician ratings, which correlate lower than coders (Pedersen et al. reported a correlation between human subjects of 0.68 for physicians and 0.78 for coders). On the other hand, coders, because of their training and skills, were more familiar than physicians to hierarchical classifications and semantic *similarity* which lead to a better correlation with the design principles of our similarity approach.

5.4 Evaluation without missing terms

Since, in the experiments reported above, some of the tests presented missing terms whose similarity estimation hampered the final correlation values in a mono-ontology setting, we ran an additional battery of tests considering only word pairs appearing in all the ontologies. In this manner, our method will always face the situation described in case 2 of section 4.2, in which it should select the best assessment from those provided by several overlapping ontologies. In addition to the benefits obtained by solving missing cases evaluated above, in the current tests, we will evaluate the absolute performance of the proposed heuristic, which takes the assessment with the highest similarity as the final result, as described in section 4.2.

In this case, only 24 of the 29 word pairs of Pedersen et al.'s benchmark and 35 of 36 word pairs of Hliaoutakis' benchmark have been found in the three ontologies. In order to quantify the differences between each individual ontology, we first computed the correlation between the similarity values obtained for each one with respect to the others. For the Hliaoutakis' benchmark the correlation between the similarity computed for each word pair in SNOMED CT with respect to the same word pairs when evaluated in MeSH was 0.636, between WordNet and SNOMED CT was 0.505 and between WordNet and MeSH was 0.630. The relatively low correlation values show a discrepancy on the way in which knowledge is represented in each ontology for this benchmark, especially for WordNet with respect to SNOMED CT, and, in consequence, a higher variance on the similarity values obtained for each ontology with respect to the same pair of words. For the Pedersen et al.'s benchmark, correlations between ontologies were much higher and constant: 0.914 for SNOMED with respect to MeSH, 0.914 for WordNet with respect to SNOMED CT and 0.904 for WordNet with respect to MeSH. In this case, the three ontologies model Perdersen et al.'s terms in a very similar manner and the differences

between ratings and the potential improvement after the heuristic is applied in a multi-ontology setting, would be less noticeable than for the Hliaoutakis' benchmark.

The re-evaluation of the same scenarios introduced in the previous section and the comparison of the similarity values with respect to human ratings results in the correlation values shown in Table 4.

Analysing the results new conclusions arise. First, those ontologies which, for the previous tests, presented a higher amount of missing concepts, now offer a higher correlation as they are not hampered by missing term pairs (e.g. 0.782 vs. 0.769 for MeSH with Pedersen et al.'s coders). In other cases in which all the word pairs were available, the correlation is lower, showing that some accurately assessed word pairs were removed from the test (e.g. 0.777 vs. 0.788 for SNOMED CT with Pedersen et al.'s coders). We see again that the combination of several ontologies leads to better results in most cases. Even though, the increase in the correlation is lower than for the first battery of tests because there are no missing concepts to solve (0.787 vs. 0.799, 0.775 vs. 0.79 and 0.798 vs. 0.817 for Pedersen et al.'s coders for SNOMED CT + WordNet, MeSH + WordNet and SNOMED CT + MeSH respectively). Only the combination of MeSH and WordNet provided a slightly lower correlation than when using MeSH alone (for the Perdersen et al.'s benchmark), even though it significantly improves WordNet's correlation alone (0.74 vs. 0.61 for WordNet and 0.749 for MeSH). In the same manner as in previous tests, the combination of all the three ontologies leads to the best results, showing the benefits of integrating the assessments from different ontologies even when missing word pairs are not considered.

Table 4: Correlation values obtained by the proposed method for Pedersen et al.'s benchmark (with 24 word pairs) for physicians', coders' and both ratings and for Hliaoutakis' benchmark (with 35 pairs) using SNOMED CT, MeSH, and WordNet.

Ontologies	Pedersen Physicians	Pedersen Coders	Pedersen Both	Hliaoutakis
SNOMED CT	0.588	0.777	0.717	0.558
MeSH	0.588	0.782	0.716	0.7496
WordNet	0.541	0.745	0.6745	0.610
SNOMED CT + WordNet	0.610	0.787	0.734	0.727
MeSH + WordNet	0.580	0.775	0.708	0.772
SNOMED CT + MeSH	0.615	0.798	0.744	0.740
SNOMED CT + MeSH + WordNet	0.638	0.812	0.760	0.786

Moreover, the correlation improvement is also coherent with the differences observed between each ontology. For example, assessments based on SNOMED CT for Hliaoutakis' benchmark improved from 0.558 to 0.727 when WordNet is also used; as

stated above, the correlation between these ontologies was 0.505 (they model Hliaoutakis' words in a noticeable different manner). On the contrary, similarity based on SNOMED CT for Perdersen et al.'s coders slightly improved from 0.777 to 0.787 when WordNet is also used; as stated above, both ontologies presented a high correlation of 0.914 indicating that Perdersen et al.'s words are modelled in a very similar way and the potential benefits of combining them would be less noticeable.

All these results show that the proposed heuristic which selects the most appropriate assessment for overlapping terms, that is, the one with the highest similarity behaves as hypothesised in section 4.2.

5.5 Comparison

Finally, in order to directly compare our method in a multi-ontology setting with the one proposed by [32], which represents, as far as we know, the most recent and complete related work, we reproduced their most complex test. In that case, the Rubenstein and Goodenough's benchmark were joined to Pedersen et al.'s and Hliaoutakis' biomedical benchmarks individually and to both of them at the same time. Note that, with regards to the human ratings, only those provided by the medical coders for the Pedersen et al.'s benchmark were used. The reason argued by Al-Mubaid and Nguyen was that medical coders' judgments were more reliable than physicians' ones because more human subjects were involved (9 coders vs. 3 physician) and because the correlation between coders were higher than between physicians (0.78 vs. 0.68).

In [32], the set of word pairs resulting from joining the two and three benchmarks were evaluated with the combination of MeSH and WordNet in first place, and with SNOMED CT and WordNet in second place. WordNet was selected as the primary ontology in all their tests. Obviously, since Rubenstein and Goodenough's terms are general words, they can only be found in WordNet, whereas the rest can be found in both WordNet and the medical ontologies in an overlapping way. It is important to note that the human ratings of the benchmarks of Pedersen et al., Hliaoutakis and Rubenstain and Goodenough have to be converted to a common scale in order to properly compute the correlation value.

In a first experiment, we used WordNet and MeSH ontologies. As stated above, 25 out of 30 pairs of Pedersen et al.'s benchmark and all the 36 pairs of Hliaoutakis's benchmark were found in MeSH. Following the experiment performed in [32], missing word terms were removed. Their correlation values in comparison with those obtained in our test are shown in Table 5.

Table 5: Correlation values obtained when joining Rubenstain and Goodenough (R&G) benchmark (65 words) with Pedersen et al.'s benchmark (with 24 pairs, only coders' ratings are considered), and Hliaoutakis' benchmark (with 36 pairs) using MeSH and WordNet.

Method	Ontologies	*R&G + Ped. (Coders)*	*R&G + Hliaoutakis*	*R&G + Ped.(Coders)+ Hliaoutakis*
Al-Mubaid & Nguyen	MESH + WordNet	0.808	0.804	0.814
Our Method	MESH + WordNet	*0.848*	*0.825*	*0.830*

In the second experiment, WordNet and SNOMED CT ontologies were used. Again, 29 out of 30 pairs of Pedersen et al.'s benchmark and 35 out of 36 pairs in Hliaoutakis' benchmark were found in SNOMED CT. Missing word pairs were removed. The results are shown in Table 6.

Table 6: Correlation values obtained when joining Rubenstain and Goodenough (R&G) benchmark (65 words) with Pedersen et al.'s benchmark (with 29 pairs, only coders' ratings are considered) and Hliaoutakis' benchmark (with 35 pairs) using SNOMED CT and WordNet.

Method	Ontologies	*R&G + Ped. (Coders)*	*R&G + Hliaoutakis*	*R&G + Ped. (Coders) + Hliaoutakis*
Al-Mubaid & Nguyen	SNOMED CT + WordNet	0.778	0.700	0.757
Our Method	SNOMED CT + WordNet	*0.850*	*0.811*	*0.816*

Analysing both tables, in all cases, our method is able to improve the results reported in [32]. This is motivated both from the higher accuracy of the underlying similarity measure in comparison with path-based ones [31] and because of the differences in the method used to select the most appropriate assessment for overlapping word pairs. In the first case, our method is able to exploit the benefits of considering additional taxonomical knowledge (all concept subsumers) instead of the path. In the latter case, the fact of relying on a *primary* ontology implies that Al-Mubaid and Nguyen's method, in many situations, omits potentially more accurate assessments which could be obtained from ontologies considered as *secondary*. On the contrary, our approach evaluates each word pair and ontology individually and homogenously, which avoids the necessity of pre-selecting a *primary* one. This exploits the benefits that each one may provide with regards to knowledge modelling.

Summarising, these results support the hypothesis introduced in section 4.2 about the benefits that our approach is able to provide not only in cases in which the pair of concepts belongs to a unique ontology, but also with multiple and overlapping ontologies.

6 Conclusions

In this paper, we studied several approaches to compute the semantic similarity in the multi-ontology scenario. Some limitations were identified, such as the fact of relying (supervision) on a predefined *primary* ontology, omitting the benefits that *secondary* ontologies may provide, the necessity to scale results in order to compare them or the complexity the integration of partial results.

To tackle these problems, we proposed a new method that avoids most of the limitations and drawbacks introduced above. The method has been designed as a set of heuristics to be applied according to the different situations that may appear when comparing concept pairs spread through several ontologies. On the contrary to [38,45,39,41] we considered all possible configurations of concept pairs, considering both the case in which overlapping knowledge is available and also the one in which partial knowledge should be integrated. Another difference is the fact that only taxonomical knowledge is needed for the similarity assessment, which is the most commonly available in ontologies. This configures a general approach that can be applied to any combination of ontologies. On the contrary to [32], our method is unsupervised, because it does not require the pre-selection of a *primary* ontology. All input ontologies are considered equally important. On the one hand, this configures a simpler scenario of only three cases instead of five. On the other hand, the adaptive heuristic proposed when overlapping knowledge is available, allows exploiting the benefits of individual ontologies for each concept pair. On the contrary to [32,45], who relied on the path to compute similarities, we exploited a similarity measure [31] that considers all taxonomical ancestors. Our ontology integration method when partial knowledge is available also enabled the evaluation of superconcepts sets rather than paths, exploiting additional taxonomical evidences of similarity.

As shown in the evaluation, these theoretical advantages have been reflected in an improved similarity accuracy when compared with related works. General purpose ontologies (WordNet) and overlapping biomedical ones (SNOMED CT and MeSH) and several standard benchmarks have been used to enable an objective comparison.

The proposed method can be potentially applied to any similarity measure fulfilling a set of requisites: to be ontology-based, to provide normalised values according to ontological dimensions and, preferably, to base the assessment solely on taxonomical knowledge. As discussed in section 4, several similarity measures fulfil these requirements. As future work, we plan to apply/adapt our method to other similarity measures in order to test its generality. Non trivial adaptations would be necessary to integrate, for example, relative taxonomical depths. Then, to test the benefits of our method in a practical setting, we plan to apply it to data mining of textual sources by means of ontology-based clustering. Because of the lack of structure of textual features and the difficulty to properly compare

textual attributes are some of the most important problems of clustering algorithms dealing with textual sources, our method could aid to improve the reliability and accuracy of the final classifications when several knowledge sources are available as background. Moreover, other ontological features (e.g. OWL features like properties, class restrictions and other non-taxonomical knowledge) could be considered. In that case, the more general notion of *semantic relatedness*, in comparison to the *semantic similarity*, which state the *taxonomic resemblance* and in which our work is framed, would be captured. Finally, considering the limitations of the existing standard benchmarks (small size and a few human experts evaluating it), we let for future work the design of a new benchmark specially focused on the multi-ontology scenario.

References

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

2. Cilibrasi RL, Vitányi PMB (2006) The Google Similarity Distance. IEEE Trans Know Data Eng 19 (3):370-383

3. Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of semantic distance. Comput Linguist 32 (1):13-47

4. Sánchez D, Isern D (2011) Automatic extraction of acronym definitions from the Web. Appl Intell 34 (2):311-327. doi:10.1007/s10489-009-0197-4

5. Sánchez D, Isern D, Millán M (2011) Content Annotation for the Semantic Web: an Automatic Web-based Approach. Knowledge and Information Systems 27 (3):393-418

6. Sánchez D, Moreno A (2008) Learning non-taxonomic relationships from web documents for domain ontology construction. Data Know Eng 63 (3):600-623

7. Sánchez D (2010) A methodology to learn ontological attributes from the Web. Data Know Eng 69 (6):573-597

8. Sánchez D, Moreno A (2008) Pattern-based automatic taxonomy learning from the Web. AI Commun 21 (1):27-48

9. Li S-T, Tsai F-C (2010) Constructing tree-based knowledge structures from text corpus. Appl Intell 33 (1):67-78

10. Iannone L, Palmisano I, Fanizzi N (2007) An algorithm based on counterfactuals for concept learning in the Semantic Web. Appl Intell 26 (2):139-159

11. Nguyen HA, Al-mubaid H (2006) New Ontology-Based Semantic Similarity Measure for the Biomedical Domain. In: IEEE Conference on Granular Computing, GrC 2006, Silicon Valley, USA, 2006. IEEE Computer Society, pp 623-628

12. Sim KM, Wong PT (2004) Toward agency and ontology for web-based information retrieval. IEEE Trans on Syst, Man, and Cybern, Part C: Appl and Rev 34 (3):257-269

13. Lee JH, Kim MH, Lee YJ (1993) Information Retrieval Based on Conceptual Distance in Is-A Hierarchies. J Doc 49 (2):188-207

14. Guarino N (1998) Formal Ontology in Information Systems. In: Guarino N (ed) 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, Trento, Italy, June 6-8 1998. Frontiers in Artificial Intelligence and Applications. IOS Press, pp 3-15

15. Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Sci Am 284 (5):34-43

16. Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts

17. Isern D, Moreno A, Sánchez D, Hajnal Á, Pedone G, Varga LZ (2011) Agent-based execution of personalised home care treatments. Appl Intell 34 (2):155-180. doi:10.1007/s10489-009-0187-6

18. Baumeister J, Reutelshoefer J, Puppe F (2011) KnowWE: a Semantic Wiki for knowledge engineering. Appl Intell 35 (3):323-344. doi:10.1007/s10489-010-0224-5

19. Eyharabide V, Amandi A (2012) Ontology-based user profile learning. Appl Intell 36 (4):857-869. doi:10.1007/s10489-011-0301-4

20. Mousavi A, Nordin MJ, Othman ZA (2012) Ontology-driven coordination model for multiagent-based mobile workforce brokering systems. Appl Intell 36 (4):768-787. doi:10.1007/s10489-011-0294-z

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

22. Li Y, Bandar Z, McLean D (2003) An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans Know Data Eng 15 (4):871-882

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

24. Rada R, Mili H, Bichnell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 9 (1):17-30

25. Lin D (1998) An Information-Theoretic Definition of Similarity. In: Shavlik J (ed) Fifteenth International Conference on Machine Learning, ICML 1998, Madison, Wisconsin, USA, July 24-27 1998. Morgan Kaufmann, pp 296-304

26. Jiang JJ, Conrath DW (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, Sep 1997. pp 19-33

27. Resnik P (1995) Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In: Mellish CS (ed) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada, August 20 - 25 1995. Morgan Kaufmann Publishers Inc., pp 448-453

28. Sánchez D, Batet M, Isern D (2011) Ontology-based Information Content computation. Knowl-Based Syst 24 (2):297-303

29. Sánchez D, Batet M, Valls A, Gibert K (2010) Ontology-driven web-based semantic similarity. J Intell Inf Syst 35 (3):383-413

30. Sánchez D, Batet M (2011) Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. J Biomed Inform 44 (5):749-759
31. Batet M, Sánchez D, Valls A (2011) An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 44 (1):118-125
32. Al-Mubaid H, Nguyen HA (2009) Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies. IEEE Trans on Syst, Man, and Cybern, Part C: Appl and Rev 39 (4):389-398. doi:10.1109/TSMCC.2009.2020689
33. Tversky A (1977) Features of similarity. Psychol Rev 84:327-352
34. Gangemi A, Pisanelli D, Steve G (1998) Ontology Integration: Experiences with Medical Terminologies. In: Guarino N (ed) Formal Ontology in Information Systems, 1998. Frontiers in Artificial Intelligence and Applications. IOS Press, pp 163-178
35. Weinstein P, Birmingham WP (1999) Comparing Concepts in Differentiated Ontologies. In: 12th Workshop on Knowledge Acquisition, Modeling and Management, KAW 1999, Banff, Alberta, Canada, 1999.
36. Mena E, Kashyap V, Sheth A (1996) OBSERVER: An Approach for Query Processing in Global Information Systems Based on Interoperation Across Pre-Existing Ontologies. In: International Conference of Cooperative Information Systems, CoopIS 1996, 1996.
37. Bergamaschi B, Castano S, Vermercati SDCd, Montanari S, Vicini M (1998) An Intelligent Approach to Information Integration. In: Guarino N (ed) Proceedings of the First International Conference Formal Ontology in Information Systems, 1998. pp 253-268
38. Rodríguez MA, Egenhofer MJ (2003) Determining semantic similarity among entity classes from different ontologies. IEEE Trans Know Data Eng 15 (2):442–456
39. Petrakis EGM, Varelas G, Hliaoutakis A, Raftopoulou P (2006) X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies. J Digit Inf Manag 4:233-237
40. Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J (2004) Swoogle: A Search and Metadata Engine for the Semantic Web. In: thirteenth ACM international conference on Information and knowledge management, CIKM 2004, Washington, D.C., USA, 2004. ACM Press, pp 652-659
41. Saruladha K, Aghila G, Bhuvaneswary A (2010) Computation of Semantic Similarity among Cross ontological Concepts for Biomedical Domain. Journal of Computing 2:111-118
42. Sánchez D, Solé-Ribalta A, Batet M, Serratosa F (2012) Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. Journal of Biomedical Informatics 45 (1):141-155
43. Al-Mubaid H, Nguyen HA (2006) A cluster-based approach for semantic similarity in the biomedical domain. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006 New York, USA, 2006. IEEE Computer Society, pp 2713–2717
44. Bollegala D, Matsuo Y, Ishizuka M (2007) WebSim: A Web-based Semantic Similarity Measure. In: 21st Annual Conference of the Japanese Society for Artificial Intelligence, JSAI 2007, Miyazaki, Japan, June 18-22 2007. pp 757-766

45. Solé-Ribalta A, Serratosa: F (2010) Exploration of the labelling Space given graph edit distance costs. Paper presented at the Workshop on Graph-based Representations in Pattern Recognition,
46. Euzenat J, Shvaiko P (2007) Ontology Matching. Springer Verlag,
47. Gómez-Pérez A, Fernández-López M, Corcho O (2004) Ontological Engineering. 2nd edn. Springer-Verlag,
48. Krumhansl C (1978) Concerning the Applicability of Geometric Models to Similarity Data: The Interrelationship between Similarity and Spatial Density. Psychol Rev 85:445-463
49. Noy NF, Musen MA (1999) SMART: Automated Support for Ontology Merging and alignment. In: B.R Gaines and B KaMAM (ed) Proceedings of the 12th Banff Workshop on Knowledge Acquisition, Modeling, and Management., Banff, Alberta, Canada, 1999. pp 1-20
50. Lambrix P, Tan H (2007) A tool for evaluating ontology alignment strategies. J Data Semant:182-202
51. Stoilos G, Stamou G, Kollias S (2005) A String Metric for Ontology Alignment. In: 4th International Semantic Web Conference, 2005. pp 624-637
52. Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. Lang Cogn Process 6 (1):1-28
53. Rubenstein H, Goodenough J (1965) Contextual correlates of synonymy. Commun ACM 8 (10):627-633
54. Pedersen T, Pakhomov S, Patwardhan S, Chute C (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40 (3):288-299
55. Hliaoutakis A (2005) Semantic Similarity Measures in the MESH Ontology and their Application to Information Retrieval on Medline. Diploma Thesis. Technical Univ. of Crete (TUC), Dept. of Electronic and Computer Engineering, Crete, Greece