A Semantic Similarity Method Based on Information Content Exploiting Multiple Ontologies

David Sánchez[1] and Montserrat Batet

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

Abstract

The quantification of the semantic similarity between terms is an important research area that configures a valuable tool for text understanding. Among the different paradigms used by related works to compute semantic similarity, in recent years, information theoretic approaches have shown promising results by computing the Information Content (IC) of concepts from the knowledge provided by ontologies. These approaches, however, are hampered by the coverage offered by the single input ontology. In this paper, we propose extending IC-based similarity measures by considering multiple ontologies in an integrated way. Several strategies are proposed according to which ontology the evaluated terms belong. Our proposal has been evaluated by means of a widely used benchmark of medical terms and MeSH and SNOMED CT as ontologies. Results show an improvement in the similarity assessment accuracy when multiple ontologies are considered.

Keywords: Information Content, Semantic Similarity, Ontologies, MeSH, SNOMED CT

[1] Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain
Tel.: +34 977559657; Fax: +34 977 559710;
E-mail: david.sanchez@urv.cat.

1. Introduction

The estimation of the semantic similarity between terms contributes to the better understanding of textual resources. As a result, it has been applied in many different tasks such as word-sense disambiguation (Resnik, 1999), document categorization or clustering (M. Batet, 2011; Cilibrasi & Vitányi, 2006; Luo, Chen, & Xiong, 2011), word spelling correction (Budanitsky & Hirst, 2006), automatic language translation (Cilibrasi & Vitányi, 2006), ontology learning (Sánchez, 2010; Sánchez & Moreno, 2008a, 2008b; Sánchez, Moreno, & Vasto, 2012), semantic annotation (Sánchez, Isern, & Millán, 2011), information extraction (Atkinson, Ferreira, & Aravena, 2009; Sánchez & Isern, 2011), information retrieval (Al-Mubaid & Nguyen, 2006; Budanitsky & Hirst, 2006) or anonymisation of textual documents (S. Martínez, Sánchez, & Valls, 2012; Sergio Martínez, Sánchez, Valls, & Batet, 2012).

Semantic similarity is understood as the degree of taxonomic proximity between terms. Similarity measures assess a numerical score that quantifies this proximity as a function of the semantic evidence observed in one or several knowledge sources. Usually, those resources consist on taxonomies and more general ontologies, which provide a formal and machine-readable way to express a shared conceptualisation by means of a unified terminology and semantic inter-relations from which semantic similarity can be assessed. In the last years, general purpose ontologies have been developed (such as WordNet) but also domain-dependant one (such as MeSH or SNOMED CT for the biomedical domain).

According to the theoretical principles and the way in which ontologies are analysed to estimate similarity, different families of methods can be identified. In a nutshell, *edge-counting* measures base the similarity assessment on the number of taxonomical links of the *minimum path* separating two concepts contained in a given ontology (Leacock & Chodorow, 1998; Li, Bandar, & McLean, 2003; Rada, Mili, Bichnell, & Blettner, 1989; Wu & Palmer, 1994). Due to their simplicity, these approaches offer a limited accuracy due to ontologies model a large amount of taxonomical knowledge that is not considered during the evaluation of the minimum path (Montserrat Batet, Sánchez, & Valls, 2011). *Feature-based* approaches estimate similarity according to the weighted sum of the amount of common and non-common features (Sánchez, Batet, Isern, & Valls, 2012). By features, authors usually consider taxonomic and non-taxonomic information modelled in an ontology, in addition to concept descriptions (e.g., glosses) retrieved from dictionaries (Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Rodríguez & Egenhofer, 2003; Tversky, 1977). Due to the additional semantic evidences

considered during the assessment, they potentially improve edge-counting approaches. However, they usually rely on non-taxonomic features that are rarely found in ontologies (Ding, et al., 2004) and require fine tuning of weighting parameters in order to integrate heterogeneous semantic evidences (Petrakis, et al., 2006).

Finally, *information content-based* approaches, which are the focus of this work, assess the similarity between concepts as a function of the Information Content (IC) that both concepts have in common in a given ontology. In the past, IC was typically computed from concept distribution in tagged textual corpora (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995). However, this introduces a dependency on corpora availability and manual tagging that hampered their accuracy and applicability due to data sparseness (Sánchez, Batet, Valls, & Gibert, 2010). To overcome this problem, in recent years, several authors have proposed ways to infer IC of concepts in an intrinsic manner from the knowledge structure modelled in an ontology (Sánchez & Batet, 2011; Sánchez, Batet, & Isern, 2011; Seco, Veale, & Hayes, 2004; Zhou, Wang, & Gu, 2008). However, the fact that intrinsic IC-based measures only rely on ontological knowledge is also a drawback because they completely depend on the degree of coverage and detail of the unique input ontology. This limitation could be overcome computing concept's IC and estimating semantic similarity from *multiple ontologies*. As stated in (Al-Mubaid & Nguyen, 2009) the exploitation of multiple ontologies provides additional knowledge that can improve the similarity estimation and solve cases in which terms are not represented in an individual ontology. This is especially interesting in domains such as the biomedical one, in which several big and detailed ontologies are available, offering overlapping and complementary knowledge about the same topics.

As it will be discussed in section 2, few works propose similarity methods supporting more than one ontology, being all of them framed in the context of edge-counting and feature-based paradigms. In this paper we present a method to extend IC-based semantic similarity measures when multiple ontologies are available. As far as we know, no similarity methods based on IC have been proposed in the past considering more than one input ontology. The method relies on a state of the art approach to compute concept's IC from an ontology in an intrinsic manner (D. Sánchez, et al., 2011). On one hand, our method permits estimating the similarity when a term or a term pair is missing in a certain ontology but it is found in another one. On the other hand, in case of overlapping knowledge (i.e. ontologies covering the same terms), our approach increases the accuracy by selecting the most reliable IC and similarity estimation from those computed from each individual ontology. The method has been evaluated by means of a widely used benchmark of biomedical terms and the above-mentioned biomedical ontologies. Results show that intrinsic IC measures are able to improve other similarity computation paradigms.

Moreover, the exploitation of several complementary and/or overlapping ontologies during the similarity assessment was able to improve the accuracy with respect to the mono-ontology scenario.

The rest of the paper is organised as follows. Section 2 introduces related works proposing methods for semantic similarity assessment from multiple ontologies. Section 3 analyses different approaches for computing the IC of a concept, focusing on ontology-based methods. Afterwards, classic IC-based similarity measures are presented. Section 4 describes our method to exploit multiple ontologies for similarity assessment, detailing the strategies proposed to tackle the problem according to which ontology the evaluated terms belong. Section 5 evaluates our approach, comparing it to a mono-ontology scenario. The final section contains the conclusions and some lines of future research.

2. Related work

Semantic similarity estimation methods supporting multiple ontologies are based on the edge-counting and feature-based paradigms.

In Rodriguez and Egenhofer (Rodríguez & Egenhofer, 2003), the similarity is computed as the weighted sum of similarities between synonym sets, features (e.g., meronyms, attributes, etc.) and neighbour concepts (those linked via semantic pointers) of evaluated terms. Petrakis et al., (Petrakis, et al., 2006) extended the previous approach relying on the matching between synonym sets and concept glosses (i.e., term definitions). They considered that two terms are similar if their synonyms and glosses and those of the concepts in their neighbourhood (following semantic relations) are lexically similar. In both approaches, when the evaluated term pair belongs to different ontologies, authors connect ontologies by a new imaginary root node that subsumes the root nodes of each ontology. Then, the similarity is computed from the resulting knowledge structure.

A problem of these approaches is the reliance on many ontological features that are rarely found in ontologies. In fact, an investigation of the structure of existing ontologies (Ding, et al., 2004) has shown that ontologies very occasionally model non-taxonomic knowledge. Another problem for Rodriguez and Egenhofer's approach is its dependency on the weighting parameters that balance the contribution of each feature. These parameters should be tuned according to the nature of the ontology and the evaluated terms. This hampers the applicability as a general purpose solution. Petrakis et al.'s method does not depend on weighting parameters,

because the maximum similarity provided by each feature alone is taken. Even though this adapts the behaviour of the measure to the characteristics of the ontology, the contribution of other features is omitted because only the maximum value is considered.

A more elaborated approach is presented in (Sánchez, Solé-Ribalta, Batet, & Serratosa, 2012). This work complements the strict matching of subsumers according to their labels with a structural similarity function that aims at discovering similar but not necessarily terminologically identical subsumers. Since only one subsumer pair is matched, the method can only be applied to path-based similarity measures.

With respect to the multi-ontology scenario, the above methods do not consider the case in which a term pair is found in several ontologies at the same time. In consequence, they omit the problem of selecting the most appropriate assessment and/or to evaluate overlapping sources of information.

A more general approach by Al-Mubaid and Nguyen (Al-Mubaid & Nguyen, 2009) propose a methodology to exploit biomedical sources (such as SNOMED CT or MeSH) using a similarity measure defined in (Al-Mubaid & Nguyen, 2006). This measure combines, in a weighted manner, the features *path length* and *common specificity* of the compared concepts. Authors quantify the common specificity of two concepts by subtracting the depth of their Least Common Subsumer (LCS) from the depth of the taxonomic branch to which they belong. In this manner, concepts at a lower level of the taxonomy are considered to be more similar those located at a higher level. In (Al-Mubaid & Nguyen, 2009) they extended this measure when multiple input ontologies are available. In their approach, the user must select a *primary* ontology (the rest are considered as *secondary*) that acts as the master in cases in which concepts belong to several ontologies. The *primary* ontology is also used as the base to normalise similarity values. Different heuristics are proposed according to which ontologies the compared concepts belong. If both concepts appear in the *primary* ontology, the similarity is computed exclusively from that source (even if they also appear in a *secondary* ontology). When concepts appear in several *secondary* ontologies, authors evaluate the degree of overlapping with respect to the *primary* ontology and the degree of taxonomic detail (granularity). The *secondary* ontology with the highest alikeness to the *primary* one is chosen. Finally, if a concept appears in an ontology and the other concept is found in another ontology, they “connect” both ontologies by finding “common nodes” (i.e., a subsumers representing the same concepts in any of the ontologies).

A problem faced by the authors is the fact that, due to their measure is based on absolute path lengths between concepts, the similarity computed for each term pair from a different ontology will lead to similarity values that cannot be directly compared. Authors 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. Both the *path* and *common specificity* features are scaled to the *primary* ontology according to difference in the depth with respect to the *primary* ontology.

Another disadvantage of this approach is the fact that similarity estimation is based on the *minimum path length* connecting concept pairs. This omits other taxonomic knowledge explicitly modelled in the ontology (i.e., other taxonomical paths). Moreover, the multi-ontology method is hampered by the fact that a *primary* ontology must be selected a priori to scale similarity values. This results in a complex casuistic to be considered during the similarity assessment. Finally, the method assumes that, in all cases, the *primary* ontology will lead to better similarity estimations than *secondary* ones. This constrains the knowledge exploitation in cases in which a particular ontology offer a better knowledge representation for the evaluated concepts.

3. Information content and semantic similarity

The Information Content (IC) of a concept states the amount of information provided by the concept when appearing in a context. In this manner, general and abstract entities present less IC when found in a discourse than more concrete and specialised ones. A proper quantification of the IC of concepts improves text understanding by enabling assessing the degree of semantic generality or concreteness of words referring to these concepts. In fact, as stated in the introduction, IC has been applied in the past to the computation of semantic similarity (Resnik, 1995) according to the amount of common information of a concept pair.

3.1. Corpora-based IC calculus methods

In classical approaches (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995) IC is computed as the inverse of the appearance probability of a concept *c* in a corpus (1).

$$IC(c) = -\log p(c) \tag{1}$$

If the corpus is large and heterogeneous enough to accurately represent concept usage at a social scale, *p(c)* will enable an accurate computation of the IC of *c*. However, textual ambiguity and data sparseness severely hamper *p(c)* estimation. First, because textual corpora contain words rather than concepts, it is necessary to disambiguate concept appearances, identifying word senses by means of manual tagging. Secondly, it is unlikely that such large and representative required corpora are available, especially for concrete domains such as biomedicine due to the sensitivity of clinical data.

It is important to note that to compute coherent values of *p(c)* from a semantic point of view, one must consider all the explicit appearances of *c* in addition to the appearances of concepts that are semantically subsumed by *c* (i.e., all its taxonomical specialisations and instances) (Resnik, 1995). For example, to estimate the IC of the concept *'neoplasm'*, all its explicit appearances should be counted along with the appearances of all its specialisations such as '*breast cancer'*, '*lung cancer'*, etc. Formally, *p(c)* is calculated as follows:

$$p(c) = \frac{\sum_{w \in W(c)} appearances(w)}{N} \quad (2)$$

, where $W(c)$ is the set of terms in the corpus whose senses are subsumed by *c* and *N* is the total number of corpus terms.

As a result, it is necessary to obtain concept specialisations from ontological resources before computing term probabilities in a corpus. The background taxonomy must be as complete as possible (i.e., it should include most of the specialisations of a specific concept) to provide reliable results.

Data sparseness, data availability dependency, and scalability problems due to the need of manually tagged corpora required to minimise language ambiguity hamper the applicability of these approaches.

3.2. Intrinsic IC computation models

To overcome these problems, in recent years, some authors have proposed computing IC in an intrinsic manner by using only the knowledge structure modelled in an ontology (Seco, et al., 2004). These works assume that the taxonomic structure of ontologies is organised in a meaningful way, according to the principles of cognitive saliency (Pirró, 2009): concepts are specialised when they must be differentiated from other ones. Concepts located at a higher level

in the taxonomy with many hyponyms or leaves (i.e., specialisations) under their taxonomic branches would have less IC than highly specialised concepts (with many hypernyms or subsumers) located on the leaves of the hierarchy.

Compared against corpora-based IC calculus, it is assumed that abstract ontological concepts (with a large set of hyponyms) are more likely to appear in a corpus than very specialised ones, because the former can be implicitly referred to in a discourse by means of all their subsumed concepts. As a result, appearance probabilities are approximated in these approaches in accordance with the number of the concept's hyponyms.

Some measures have been published in recent years proposing intrinsic IC computation models based on the structural principles of knowledge discussed above (Seco, et al., 2004; Zhou, et al., 2008). In a recent work (D. Sánchez, et al., 2011), we improved them by incorporating additional semantic evidence extracted from the input ontology into the assessment. We proposed estimating *p(c)* as the ratio between the number of *leaves* on the taxonomical hierarchy under the concept *c* (as a measure of *c*'s generality) and the number of taxonomical *subsumers* above *c* including itself (as a measure of *c*'s concreteness) (3). It is important to note that in case of multiple inheritance all the ancestors are considered. Formally:

$$IC(c) = -\log p(c) \cong -\log\left(\frac{\frac{|\, leaves(c)\,|}{|\, subsumers(c)\,|} + 1}{max\_leaves + 1}\right) \quad (3)$$

The above ratio has been normalised by the least informative concept (i.e., the root of the taxonomy), for which the number of leaves is the total amount of leaves in the taxonomy (*max_leaves*) and the number of subsumers including itself is 1. To produce values in the range 0..1 (i.e., in the same range as the original probability) and avoid *log(0)* values, 1 is added to the numerator and denominator.

As discussed in (D. Sánchez, et al., 2011) this approach represents an improvement to previous ones (Seco, et al., 2004; Zhou, et al., 2008) in that it can differentiate concepts with the same number of hyponyms/leaves but different degrees of concreteness (expressed by the number of subsumers that normalises the numerator). It can also consider the additional knowledge modelled by means of multiple inheritance relationships. Finally, it prevents the dependence on the granularity and detail of the inner taxonomical structure by relying on taxonomic *leaves* rather than complete sets of hyponyms.

3.3. IC-based semantic similarity

Accurate quantification of the IC of concepts permits the estimation of their semantic similarity as a function of their shared information.

In Resnik's (Resnik, 1995) seminal work, he proposed evaluating the IC of the Least Common Subsumer of the compared concepts ($LCS(c_1,c_2)$) as the representative of this shared information. The LCS of a pair of terms in a taxonomy/ontology is the most specific common ancestor that subsumes them, found in the taxonomy to which they belong. If the two concepts are not taxonomically connected and the LCS does not exist, they are considered maximally different. Otherwise their semantic similarity is computed as the amount of IC provided by the LCS (4).

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \quad (4)$$

With Resnik's metric, any pair of concepts with the same LCS will result in exactly the same similarity value. To better differentiate concepts, both Lin (Lin, 1998) and Jiang and Conrath (Jiang & Conrath, 1997) also consider the IC of the compared terms into the equations.
Lin measures the similarity as the ratio between the common information between concepts (i.e., IC(LCS)) and the information needed to fully describe them (i.e., the IC of each concept alone) (5).

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(LCS(c_1, c_2))}{(IC(c_1) + IC(c_2))} \quad (5)$$

Jiang and Conrath proposed calculating the concept distance (the opposite of similarity) as the difference between the IC of each concept and the IC of their LCS (6).

$$dis_{j\&c}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2 \times IC(LCS(c_1, c_2)) \quad (6)$$

4. Extending IC-based measures to multiple ontologies

The availability of several knowledge sources can potentially aid the similarity assessment in cases in which a concept or a concept pair is missing in an ontology but found in another, or in situations in which ontologies overlap. In this section we present a method to extend IC-based similarity measures to take profit from multiple input ontologies.

Regarding IC-based similarity estimation, as shown in the previous section, the key point to compare a pair of concepts is to retrieve their LCS. According to the ontology to which the compared concepts belong, we can distinguish three cases. In the first one, both concepts appear in a unique ontology, so that the LCS can be retrieved from that one, and the similarity can be computed in the same manner as in the mono-ontology scenario. Other cases in which term pairs appear, at the same time, in several ontologies or in which each term belongs to a distinct ontology require the definition of specific strategies to solve them.

4.1. Computing IC-based similarity from multiple overlapping ontologies

If both concepts appear in several ontologies, each one modelling the same knowledge in a different but overlapping way, several LCSs can be retrieved. In this case, it is necessary to decide which LCS is the most suitable to compute inter-concept similarity. Considering the implications of the ontology engineering process and the way in which humans assess similarity, two hypotheses can be enounced.

First, ontological knowledge modelling is the result of a manual an explicit engineering process performed by domain experts. In other words, the fact that a concrete LCS is represented *near* of a pair of concepts (e.g., the LCS of *flu* and *bronchitis* is *respiratory disease*) is the result of a decision taken by the human expert stating a high degree of commonalty between concepts' meaning. However, due to the coverage limitations and the bottleneck that characterise manual knowledge modelling processes, if a LCS is missing or appears in a very abstract level in the hierarchy (e.g., the LCS of *flu* and *bronchitis* is *condition*), one cannot ensure if this is an implicit indication of semantic disjunction between the compared concepts or the result of partial or incomplete knowledge modelling. Summarising, we hypothesise that the fact that a *specific* LCS (from a taxonomical perspective) is available in an ontology for a pair of concepts is more important when computing their similarity than its absence or the presence of a *general* one.

The second hypothesis regards human perception of similarity. As demonstrated in psychological studies (Tversky, 1977), humans pay more attention to common than to differential features of compared entities.

As a result of these arguments, given a pair of concepts belonging to different ontologies, we consider the *most specific* LCS from those retrieved from the overlapping set of ontologies. In

terms of IC, this corresponds to the LCS with the maximum IC value (i.e., the maximum specificity) (7).

Definition 1. Given a pair of compared concepts c_1 and c_2 and a set of ontologies O to which they belong, the selected *LCS* is:

$$LCS(c_1, c_2) = \underset{\forall o \in O \mid c_1, c_2 \in o}{\arg\max} (IC(LCS_o(c_1, c_2)) \quad (7)$$

,where $LCS_o(c_1,c_2)$ is the *LCS* between c_1 and c_2 for the ontology $o \in O$.

The selected LCS and its corresponding IC value can be directly applied in conjunction with concept's IC (computed as stated in eq. 3) over the IC-based similarity measures described above.

It is important to note that, for a set of concept pairs belonging to several ontologies, similarity values for each pair may be assessed from IC values computed from different ontologies. This is the result of selecting, for each concept and concept pair's LCS, the IC assessment that better represents concepts' meaning and that more accurately differentiates them from other modelled entities. As a result, the local benefits that each ontology may provide (regarding accurate knowledge modelling of a particular concept) are exploited during the similarity estimation. This contrast to other approaches (Al-Mubaid & Nguyen, 2009) relying on a *primary* ontology, which prioritise a pre-selected ontology over *secondary* ones, imposing a hard dependency. The fact that all input ontologies are equally considered in our approach also simplifies the multi-ontology casuistry proposed by related works (Al-Mubaid & Nguyen, 2009).

4.2. Computing IC-based similarity from multiple disjoint ontologies

This is the case in which each concept of the compared pair belongs to a different ontology (e.g., $c_1 \in o_1, c_2 \in o_2 \mid o_1 \neq o_2$).

As stated in (Rodríguez & Egenhofer, 2003), similarity estimation from different ontologies can only be achieved if they share some components. In some related works (Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003), the two ontologies are simply connected by creating a new node (called *anything*) which is a direct ancestor of their roots. This avoids the problem of knowledge integration but poorly captures possible commonalities between ontologies. As stated in section 2, other authors (Al-Mubaid & Nguyen, 2009) base their proposal in the differentiation between *primary* and *secondary* ontologies, connecting them by joining all nodes with the same textual label. These nodes are called *bridges*. Then, the LCS of the pair of concepts in two different

ontologies is redefined as the LCS of the concept belonging to the *primary* ontology and each *bridge* node. The path length is then computed via the LCS and the *bridge* node. Due to the path length depends on the granularity and size of a concrete ontology, it is normalised by measuring 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.

We follow a similar principle to tackle this situation, but considering the two ontologies equally important. We retrieve the set of subsumers for each of the compared concepts (each one belonging to a different ontology). Then, both sets are compared to find *equivalent* subsumers. Equivalent subsumers are those with the same textual label considering, if available, synonym sets (which is the case of MeSH and SNOMED CT, for example). As a result, the two ontologies can be "connected" by a set of equivalent and the LCS for the concept pair can be retrieved as the *least common equivalent subsumer*, similarly to the mono-ontology scenario. In the worst case, when no terminologically equivalent ancestors are found, we consider the root nodes of both ontologies as equivalent and, hence, they will correspond to the LCS.

Definition 2. The *LCS* between concepts c_1 and c_2, where $c_1 \in o_1, c_2 \in o_2 \mid o_1 \neq o_2$, is obtained as follows*:*

$$LCS(c_1, c_2) = Least\_Common\_Equivalent\_Subsumer(c_1, c_2) \quad (8)$$

, where the *Least_Common_Equivalent_Subsumer* is a function that terminologically compares all the subsumers of c_1 in o_1 and c_2 in o_2, and selects the most specific common one.

In any case, the IC value of the retrieved LCS (than will be necessarily common to both ontologies) will be different when computing it from an ontology than from the other.

Considering the hypotheses discussed in section 4.1 regarding the convenience of selecting the maximum IC value in cases of overlapping knowledge, we followed the same strategy.

Definition 3. The *IC* of the *LCS* selected for the concepts c_1 and c_2, where $c_1 \in o_1, c_2 \in o_2 \mid o_1 \neq o_2$, is computed as*:*

$$IC(LCS(c_1, c_2)) = \max_{o \in \{o_1, o_2\}} IC_o(LCS(c_1, c_2)) \quad (9)$$

This case can be generalised if c_1 and/or c_2 *individually* belong to several ontologies (e.g. c_1 belongs to o_1 and o_3, and c_2 belongs to o_2 and o_4). Following the same strategy, the ontology alignment process, the LCS selection and its IC calculus are executed for each combination of ontology pairs (e.g. o_1 - o_2, o_1 - o_4, o_3 - o_2 and o_3 - o_4). Again, the LCS with the maximum IC value is taken as the final result.

It is important to note that, on the contrary to related works dealing with the multi-ontology scenario by means of absolute similarity values (like path lengths (Al-Mubaid & Nguyen, 2009)), the comparison of IC values computed from different ontologies is not problematic. Given that our IC measure (eq. 3) is a *normalised ratio* between the degree of *generality* (i.e., number of leaves) and *concreteness* (i.e., number of subsumers) of the evaluated term, we avoid depending on the size, taxonomic detail or granularity of the exploited knowledge structures. In other words, it is not necessary to scale resulting values to a common factor (e.g., the size or granularity of a *primary* predefined ontology like in (Al-Mubaid & Nguyen, 2009)). This enables a coherent comparison and ranking of IC values computed from heterogeneous ontological sources.

5. Evaluation

In order to evaluate the benefits that multiple ontologies bring to similarity assessments, we have applied IC-based measures introduced in section 3.3 to several mono and multi-ontology scenarios using, in this last case, the strategies proposed in section 4.

To enable the multi-ontology setting, we have selected a domain in which several detailed and partially overlapping ontologies are available: biomedicine. In this context, SNOMED CT and MeSH knowledge sources have been used as background ontologies. They are characterised by a high level of detail, classifying concepts in different but overlapping hierarchies. The *Systematized Nomenclature of Medicine, Clinical Terms* (SNOMED CT) (Spackman, 2004) covers more than 300,000 concepts organised into 18 overlapping hierarchies. The *Medical Subject Headings* (MeSH) (Nelson, Johnston, & Humphreys, 2001) offers a hierarchy of around 25,000 medical and biological terms organised in 16 categories aimed to classify medical literature.

To provide an objective evaluation in this domain, we have taken a widely used benchmark (Al-Mubaid & Nguyen, 2009; Al-Mubaid & Nguyen, 2006; Montserrat Batet, et al., 2011; Pedersen,

Pakhomov, Patwardhan, & Chute, 2007) of medical term pairs whose similarity has been assessed by human experts. This benchmark was created by Pedersen et al. (Pedersen, et al., 2007) in collaboration with experts of the Mayo Clinic and consists of 29 words pairs. The similarity between each pair was assessed by 12 medical experts in a scale between 1 and 4. The average value of individual ratings is used in our case to compare the accuracy of our approach, according to its *correlation* against averaged human ratings. This permits an objective evaluation of similarity measures exploiting different ontologies as knowledge sources and using different strategies to assess the similarity.

5.1. Evaluation in a mono-ontology scenario

First, we show the benefits that IC-based measures computed in an intrinsic manner (IC computed as in eq. 3) bring with respect to their corpora-based counterparts (IC computed as in eq. 1), and with respect to other similarity estimation paradigms (edge-counting and feature-based). We compiled previously published results reported by related works when evaluating corpora-based IC measures (rows 5 to 7 in Table 1) and edge-counting/feature-based measures (rows 1 to 4 in Table 1) using Perdersen et al.'s benchmark and SNOMED CT as ontology. Reported corpora-based results rely on the semi-structured Mayo Clinical corpus of Medical Notes (Pedersen, et al., 2007) to compute term appearance frequencies. These results are compared against the same IC-based measures when computing concept's IC in an intrinsic manner as stated in eq. 3 (rows 8 to 10 in Table 1).

Table 1. Correlation values obtained for each measure against human ratings of Pedersen et al.'s benchmark using SNOMED CT as ontology.

Measure	Type	Evaluated in	Correlation
Rada (Rada, et al., 1989)	Edge-counting	(Pedersen, et al., 2007)	0.48
Wu and Palmer (Wu & Palmer, 1994)	Edge-counting	(Al-Mubaid & Nguyen, 2006)	0.30*
Leacock and Chodorow (Leacock & Chodorow, 1998)	Edge-counting	(Pedersen, et al., 2007)	0.47
Al-Mubaid and Ngyugen (Al-Mubaid & Nguyen, 2006)	Features+Edge-counting	(Al-Mubaid & Nguyen, 2006)	0.66*
Resnik (Resnik, 1995)	Corpora-based IC	(Pedersen, et al., 2007)	0.55
Lin (Lin, 1998)	Corpora-based IC	(Pedersen, et al., 2007)	0.69
Jiang and Conrath (Jiang & Conrath, 1997)	Corpora-based IC	(Pedersen, et al., 2007)	0.55

Resnik (Resnik, 1995)	Intrinsic IC (eq. 3)	This work	0.741
Lin (Lin, 1998)	Intrinsic IC (eq. 3)	This work	0.762
Jiang and Conrath (Jiang & Conrath, 1997)	Intrinsic IC (eq. 3)	This work	0.743

** Only 9 out of 12 experts' ratings were considered to maximise inter-human agreement.*

Results show that IC measures based on intrinsic IC calculus obtain higher correlation values than those based on corpora (0.55-0.69 vs. 0.74-0.76). This fact is very convenient because pure ontology-based similarity computation paradigms avoid depending on the availability of corpora and manual data pre-processing. On the contrary, intrinsic IC calculus models are efficient and easily applicable to any knowledge source represented in an ontological way.

On the contrary to other similarity estimation paradigms based on non-taxonomical knowledge (like glosses or meronyms (Petrakis, et al., 2006; Rodríguez & Egenhofer, 2003)) and relying on weighting factors to integrate the contributions of different semantic features (like in (Al-Mubaid & Nguyen, 2009)), intrinsic IC models only require taxonomical knowledge (which is common to any ontology, and the most structure-building component (Ding, et al., 2004; Rada, et al., 1989)). Moreover, their accuracy is higher in all cases (0.66 vs 0.74-76).

Compared against measures relying only on the minimum path length between concepts (which are also based on taxonomical knowledge), the differences are even higher. Edge-counting measures offer very limited accuracy (0.3-0.48). This shows that the minimum path between concepts poorly captures the semantics explicitly modelled in taxonomies. In fact, in complex ontologies such as SNOMED CT, several paths exist between concepts due to multiple taxonomical inheritance, a fact that is ignored by edge-counting measures.

5.2. Evaluation in a multi-ontology scenario

Regarding the multi-ontology scenario we configured two batteries of tests using the three IC-based measures and computing IC as in eq. 3.

In the first one, all the 29 words pairs of the benchmark were compared using SNOMED CT and MeSH individually and both at the same time (applying the strategies proposed in section 4). Note that all these term pairs are contained in SNOMED CT, whereas only 25 of them are found in MeSH. As a result, when evaluating term pairs over MeSH there will be some situations in which a term is missing in this ontology but found in SNOMED CT. The multi-ontology approach proposed in this paper will solve these cases as proposed in section 4. To introduce a proper penalisation in the correlation when missing word pairs appear in a mono-

ontology scenario (enabling a fair comparison against the multi-ontology setting), the similarity value of a missing word pair is replaced in the mono-ontology scenario by minimal similarity value. Correlations for this test are shown in Table 2.

Table 2. Correlation values obtained for intrinsic IC-based measures against human ratings of Pedersen et al.'s benchmark (29 word pairs) in mono and multi-ontology scenarios.

Measure	Ontologies	Correlation
Resnik	SNOMED CT	0.741
Resnik	MeSH	0.699
Resnik	**SNOMED CT+MeSH**	**0.754**
Lin	SNOMED CT	0.762
Lin	MeSH	0.705
Lin	**SNOMED CT+MeSH**	**0.762**
Jiang and Conrath	SNOMED CT	0.743
Jiang and Conrath	MeSH	0.671
Jiang and Conrath	**SNOMED CT+MeSH**	**0.744**

In all cases, we can see that the combination of several ontologies leaded to an equal or even higher correlation against human experts than when using ontologies individually. It is interesting to observe that correlations obtained from MeSH are lower with respect to those obtained from SNOMED CT. This can be motivated by the lower taxonomical detail offered by MeSH (containing around 25,000 terms) with respect to the much larger structure provided by SNOMED CT (with more than 300,000 entities). Due to the coarser granularity of MeSH's taxonomic structure, concepts tend to be less differentiated (by means of their taxonomical ancestors and hyponyms) than in SNOMED CT. Even though, when combining both ontologies, we observe in some cases a slight increase in the correlation with respect to SNOMED CT (e.g. 0.75 vs. 0.74 for Resnik measure). This suggests that, even though SNOMED CT offers, in general, more taxonomical detail and better differentiated concepts, MeSH can also contribute in punctual situations, providing better semantic similarity assessments. The strategies proposed in section 4.1 contribute to identify these situations, selecting the best assessment from those individually computed from each ontology. The increase in correlation of the multi-ontology scenario with respect to MeSH is more noticeable (0.67-0.7 vs. 0.74-76). This is motivated both by the resolving of missing values (4 word pairs in the case of MeSH) and thanks to the selection of the most accurate assessment (provided in most cases by SNOMED CT) when overlapping knowledge is available.

Missing terms affected the similarity (and the correlation) obtained from MeSH in the above test. The second battery of experiments omits word pairs missing in MeSH ontology. As a

result, only 25 word pairs have been evaluated for both MeSH and SNOMED CT. In this manner, our method will always face the situation described in section 4.1, in which it should select the best LCS from those computed/extracted from overlapping ontologies. In this manner, the contribution of the proposed strategies can be better quantified. Correlation values for this configuration are shown in Table 3.

Table 1. Correlation values obtained for intrinsic IC-based measures against human ratings of Pedersen et al.'s benchmark without considering missing terms (25 word pairs) in mono and multi-ontology scenarios.

Measure	**Ontologies**	**Correlation**
Resnik	SNOMED CT	0.736
Resnik	MeSH	0.700
Resnik	**SNOMED CT+MeSH**	**0.750**
Lin	SNOMED CT	0.758
Lin	MeSH	0.707
Lin	**SNOMED CT+MeSH**	**0.758**
Jiang and Conrath	SNOMED CT	0.742
Jiang and Conrath	MeSH	0.706
Jiang and Conrath	**SNOMED CT+MeSH**	**0.760**

As a result of removing missing terms from the test, correlation values obtained from MeSH tended to increase (e.g. 0.67 vs. 0.71 for Jiang and Conrath measure). Due to this test focuses only on overlapping knowledge, the increase in the accuracy with respect to the mono-ontology setting tends to be more noticeable for some measures (e.g. 0.74 vs. 0.76 for the Jiang and Conrath measures). These results support the suitability of the proposed strategies and the assumption that a higher IC (computed in an intrinsic manner from an ontology) better captures concepts' semantic for similarity assessments.

6. Conclusions

The fact that multiple input ontologies are available permits: *i)* to compute the similarity of concepts missing in one ontology but present in another, and *ii)* to select the most accurate estimation from those computed from different ontologies in case of overlapping knowledge (i.e., concepts belonging to several ontologies at the same time). The former case improves the recall of the similarity estimation and avoids depending on the coverage of an individual source, a serious limitation of previous ontology-based approaches (Al-Mubaid & Nguyen, 2009). The latter exploits the punctual benefits offered by each individual ontology regarding the knowledge representation adequacy in cases in which overlapping knowledge is available. In

this paper, several strategies are proposed to enable these advantages, extending the IC-based similarity estimation to the multi-ontology scenario.

The evaluation, based on a widely used benchmark and several standard biomedical ontologies, sustained the hypotheses of our approach. In all cases, the accuracy resulting from the multi-ontology scenario equalled or even increased the best accuracy observed in a mono-ontology setting.

As future work, we plan to apply our method to other domains and ontologies. The discovery of an equivalent LCS between different ontologies can also be improved by complementing the strict terminological matching with an analysis of the structural resemblance of different ontologies.

Acknowledgements

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

References

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

Al-Mubaid, H., & Nguyen, H. A. (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* (pp. 2713–2717). New York, USA: IEEE Computer Society.

Atkinson, J., Ferreira, A., & Aravena, E. (2009). Discovering implicit intention-level knowledge from natural-language texts. *Knowledge-Based Systems, 22*, 502-508.

Batet, M. (2011). Ontology-based Semantic Clustering. *AI Communications, 24*, 291-292.

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

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics, 32*, 13-47.

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

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., 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* (pp. 652-659). Washington, D.C., USA: ACM Press.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference on Research in Computational Linguistics, ROCLING X* (pp. 19-33). Taipei, Taiwan.

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

Li, Y., Bandar, Z., & McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering, 15*, 871-882.

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

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications, 38*, 12708-12716.

Martínez, S., Sánchez, D., & Valls, A. (2012). Semantic adaptive microaggregation of categorical microdata. *Computers & Security, 31*, 653-672.

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

Nelson, S. J., Johnston, D., & Humphreys, B. L. (2001). Relationships in Medical Subject Headings. In *Relationships in the Organization of Knowledge* (pp. 171-184): K.A. Publishers.

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

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

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

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

Resnik, P. (1995). Using Information Content to Evalutate Semantic Similarity in a Taxonomy. In C. S. Mellish (Ed.), *14th International Joint Conference on Artificial Intelligence, IJCAI 1995* (Vol. 1, pp. 448-453). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.

Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11*, 95-130.

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

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

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

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

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

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

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

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

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

Sánchez, D., & Moreno, A. (2008b). Pattern-based automatic taxonomy learning from the Web. *AI Communications, 21*, 27-48.

Sánchez, D., Moreno, A., & Vasto, L. D. (2012). Learning relation axioms from text: An automatic Web-based approach. *Expert Systems with Applications, 39*, 5792-5805.

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*, 141-155

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

Spackman, K. (2004). SNOMED CT milestones: endorsements are added to already-impressive standards credentials. *Healthcare Informatics, 21*, 54-56.

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

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

Zhou, Z., Wang, Y., & Gu, J. (2008). A New Model of Information Content for Semantic Similarity in WordNet. In S. S. Yau, C. Lee & Y.-C. Chung (Eds.), *Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008* (pp. 85-89). Sanya, Hainan Island, China: IEEE Computer Society.