An information theoretic approach to improve semantic similarity assessments across multiple ontologies

Montserrat Batet[1,*], Sébastien Harispe[2], Sylvie Ranwez[2], David Sánchez[1], Vincent Ranwez[3]

[1] *Department d'Enginyeria Informàtica i Matemàtiques,*

Univeritat Rovira i Virgili, Av. Paisos Catalans, 26, 43007, Tarragona, Spain

[2] *LGI2P/ENSMA Research Centre, Site EERIE,*

Parc scientifique G. Besse, 30035 Nîmes cedex 1, France.

[3]*Montpellier SupAgro, UMR AGAP,*

2 place Pierre Viala, 34060 Montpellier cedex 1, France.

[*] 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: montserrat.batet@urv.cat.

Abstract

Semantic similarity has become, in recent years, the backbone of numerous knowledge-based applications dealing with textual data. From the different methods and paradigms proposed to assess semantic similarity, ontology-based measures and, more specifically, those based on quantifying the Information Content (IC) of concepts are the most widespread solutions due to their high accuracy. However, these measures were designed to exploit a single ontology. They thus cannot be leveraged in many contexts in which multiple knowledge bases are considered. In this paper, we propose a new approach to achieve accurate IC-based similarity assessments for concept pairs spread throughout several ontologies. Based on Information Theory, our method defines a strategy to accurately measure the degree of commonality between concepts belonging to different ontologies—this is the cornerstone for estimating their semantic similarity. Our approach therefore enables classic IC-based measures to be directly applied in a multiple ontology setting. An empirical evaluation, based on well-established benchmarks and ontologies related to the biomedical domain, illustrates the accuracy of our approach, and demonstrates that similarity estimations provided by our approach are significantly more correlated with human ratings of similarity than those obtained via related works.

Keywords: semantic similarity, Information Theory, ontologies, MeSH, SNOMED-CT

1 Introduction

Semantic similarity is a pillar of text understanding since it quantifies the degree of resemblance between the meanings of textual terms. In recent years, due to the marked increase in electronically available textual data, considerable effort has been focused on defining semantic measures, which have been extensively applied in various contexts such as information retrieval [54], information extraction [56], word sense disambiguation [32], data clustering [4, 27], data privacy [6, 45-47] and biomedicine (e.g. protein classification and interaction [19, 57], chemical entity identification [18], identification of the communalities between brain data and linguistic data [15], etc.) to cite a few. Different knowledge sources have been used to facilitate the semantic similarity calculus, including: measures relying solely on textual corpora to estimate similarity from the degree of co-occurrence of terms [11], and measures involving structured knowledge bases such as ontologies [36, 37, 44], whereby the similarity calculus is based on the analysis of semantic relationships modelled between concepts. Compared to corpora-based approaches, ontology-based measures have a dual benefit: concept meanings can be unambiguously retrieved from ontologies and similarities can be assessed from structured knowledge that has been explicitly formalised by human experts.

A plethora of ontology-based measures have been proposed to estimate the similarity of two concepts belonging to a *single* ontology. They can be roughly classified as:

i) *Edge-based measures*. These measures consider an ontology as a directed graph in which concepts are interrelated by means of semantic links. They estimate similarity according to the number of semantic links separating concept pairs [24, 36, 58]. However, the fact that they only consider the minimum number of taxonomic links between concept pairs limits their accuracy, because much of the knowledge modelled in the ontology is omitted (e.g. taxonomic links resulting from multiple inheritance).

ii) *Feature-based measures*. These methods try to overcome the limitations of edge-based measures by building sets of features that describe the concepts. They estimate similarity as a function of the amount of overlapping and non-overlapping knowledge features (e.g. taxonomic ancestors, concept descriptions, etc.) between the compared concepts [34, 44].

iii) *Information Content-based measures*. These measures complement the taxonomical knowledge provided by an ontology with the quantification of the amount of information (i.e. Information Content, IC) that concepts have in common [22, 25, 37]. The IC of a concept is usually computed as the inverse of the probability of occurrence of that concept in a given corpus. However, in order to avoid textual ambiguity and data sparseness, this IC calculus requires from a large, heterogeneous and tagged corpora, which is not usually available. Because of these limitations, some authors intrinsically derive IC values from an ontology according the number of taxonomic descendants and/or ancestors of concepts [41, 43, 52, 59]. These latter approaches, which constitute the focus of our work, have proved successful in mimicking human judgements of semantic similarity [17, 40, 43].

Numerous ontologies are currently available due to the widespread adoption of the Semantic Web paradigm. However, because of the complicated maintenance of large ontologies, knowledge modellers tend to spread knowledge through multiple interlinked domain ontologies (e.g. BioPortal, a portal dedicated to biomedical ontologies, provides up to 200 ontologies [30]). An increasing number of applications thus require multiple knowledge bases and ontologies to be taken into account. As an example, documents annotated by multiple ontologies may need to be queried in pluridisciplinary projects (e.g. biodiversity protection requires simultaneous consideration of economical, geographical and biological information) [21]. In order to use an information retrieval system for this task (e.g. OBIRS [54]), the semantic similarity between the query and the document annotations must be evaluated because associated knowledge may be

spread in different ontologies. However, due to their monolithic design principles, classic similarity measures cannot be applied when concepts belong to *different* ontologies [2, 7].

Such multi-ontology scenarios are common when dealing with cross-domain data (e.g. social and computer sciences) and can also occur in specific fields, such as biomedicine, in which concepts are modelled in several knowledge sources, e.g. SNOMED-CT (Systemized Nomenclature of Medical Clinical Terms) [53] and MeSH (Medical Subject Headings) [29] are knowledge bases with different scopes and purposes, but both model biomedical concepts. Indeed, many ontologies share some common knowledge even if they were initially designed for unrelated applications. They hence model complementary and overlapping aspects of a complex reality that has been split for convenience or historic reasons. Their overlaps are cornerstones for designing the multi-ontology semantic measures required in these scenarios.

Similarity measures coping with multiple ontologies have been seldom considered in the literature [2, 7, 38, 42, 51]. In the context of IC-based measures, the identification of the *Most Informative Common Ancestor* (or MICA), which represents the commonality between compared concepts, is essential for similarity assessments. Existing works based on IC [42, 51] retrieve the MICA of a pair of concepts belonging to different ontologies by looking for equivalences of concept ancestors sharing the same linguistic labels (i.e. terminological matchings). These approaches are hampered by the fact that ontologies rarely model concepts in the same way or refer to them using the same label (e.g. due to synonymy) [50]. Indeed, in many cases, they either select too abstract ancestors as MICA or they cannot discover any equivalence at all because they miss suitable concepts sharing similar meanings but referred with different labels (e.g. *cancer/neoplasm*). In both cases, the concept similarity is largely underestimated.

In this paper, we propose a method that overcomes the limitations of strict terminological matching between concept ancestors. Based on the Information Theory and solely exploiting ontological knowledge, our approach measures the degree of semantic equivalence between concept ancestors belonging to different ontologies to select a MICA that is more suitable than that obtained via terminological matchings. An empirical evaluation carried on several widespread similarity benchmarks, ontologies and classic IC-based measures shows that, by means of our method, similarity assessments obtained in a multi-ontology setting are more accurate than those obtained from related works.

The rest of the paper is organised as follows. Section 2 introduces similarity measures based on IC and details different IC computational models. Section 3 reviews related works on similarity measures handling multiple ontologies and discusses their main limitations. Section 4 presents and formalises our approach. Section 5 details the evaluation protocol and discusses the results obtained for several benchmarks, ontologies and measures. The last section gives the conclusions of this work.

2. Semantic similarity based on Information Content

In the Information Theory context, the informativeness of a concept *c*, i.e. its IC, is measured as the inverse of its probability of occurrence [37].

$$IC(c) = -\log p(c) \quad (1)$$

Given that concepts are assumed to occur when any of their specialisations occur (e.g. *dog* –> *mammal*), general concepts are assumed to provide less information than more specialised ones, since the former are more likely to arise in a discourse. By relying on the notion of IC**,** several IC-based similarity measures have been developed.

2.1 IC-based similarity measures

Similarity measures rely on functions that quantify the commonalities and differences of the compared concepts [55]. For IC-based measures, the commonality of two concepts c_1 and c_2 belonging to an ontology is assessed according to the informativeness of their *Most Informative Common Ancestor* (*MICA*(c_1,c_2)), which is an estimator of the information shared between both concepts. Consequently, the informativeness of this MICA constitutes the core of the semantic similarity assessment [37].

The first similarity measure relying on this principle was proposed by Resnik [37], who defined similarity according to the IC of the MICA.

$$sim_{res}(c_1, c_2) = IC(MICA(c_1, c_2)) \quad (2)$$

With this measure, the IC increases as the MICA becomes more specific and hence the compared concepts will be more similar. Note that, according to Resnik's measure, all pairs of concepts sharing the same MICA will have the same similarity. Hence, different concept pairs may result in identical similarities, since their MICA is the same., Other authors also considered the IC of each concept in the similarity assessment to achieve a more precise differentiation between concept similarities.

As an example, Lin proposed to measure the similarity of two concepts as the ratio between the information that those concepts have in common (i.e. *IC*(*MICA*(c_1,c_2))) and the information they encompass [25].

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(MICA(c_1, c_2))}{(IC(c_1) + IC(c_2))} \quad (3)$$

In this manner, the informativeness of concepts is considered during the similarity assessment. Hence, if two pairs of concepts share the same MICA, the most general pair (i.e. the one with the lowest IC values) will have a higher similarity than the most specific pair (i.e. the one with the highest ICs).

According to the same principle, Jiang and Conrath proposed a distance measure (the inverse of similarity) based on the difference between the sum of the individual ICs of the two concepts and the IC of their MICA [22].

$$dis_{j\&c}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2 \times IC(MICA(c_1, c_2)) \quad (4)$$

From these measures, we can conclude that IC-based similarity depends on:

1. The ability to retrieve an appropriate MICA that subsumes the meaning of the compared concepts.
2. The ability to estimate the IC of concepts involved in the similarity assessment.

Even though both tasks are straightforward when concepts belong to the same ontology, they represent a challenge when each belongs to a different ontology. In this latter case, a suitable pair of concept subsumers between the two ontologies should be discovered to act as a *virtual* MICA. Moreover, a valid IC value should be attributed to this MICA-like pair. These are the problems tackled in this paper with the aim of coming up with a meaningful IC-based similarity between concepts belonging to different ontologies.

Related works on this task are discussed in section 3. The next section reviews IC computation models upon which the above measures rely.

2.2 Information Content computation models

IC calculus depends on accurate assessment of the probability of appearance of concepts. If this probability faithfully represents the usage of concepts at a social scale, then the IC values will accurately reflect the concept semantics as used by human beings, which is a prerequisite for meaningful similarity assessments [37].

To capture this social dimension, probabilities are usually computed from corpora, by counting the number of times a concept appears [37]. This IC assessment approach is usually considered as *extrinsic* since it incorporates semantic evidence external to the ontology. The probability calculus will be precise if the corpora are large and heterogeneous enough to be considered as a representative sample of social usage of concepts. However, textual ambiguity and data sparseness severely hamper such probability calculus.

Since textual corpora contain words rather than concepts, textual terms should be disambiguated in order to coherently compute the number of concept appearances. This is usually done by manually tagging the concept (i.e. sense) a word is referring to within a context. In Resnik's approach, the probability of a concept *c* is computed based on *all* appearances of *c*, which can be explicit (i.e. *c* appears) or implicit (i.e. an hyponym of *c* appears) [37]:

$$p(c) = \frac{\sum_{c_h \in hypo(c)} appearances(c_h)}{N} \quad (5)$$

where *hypo*(*c*) is the set of hyponyms of *c* (i.e. the set of concepts subsumed by *c*) including itself, and *N* is the total number of concept appearances within this corpus. The monotonically increase in the probability is achieved as one moves up in the concept taxonomy by taking hyponym appearances into account, which in turn ensures that a concept consistently has an IC lower than those of its hyponyms [37].

To provide robust IC estimations, tagged corpora should be large enough to avoid data sparseness. However, since manual text tagging is time consuming, suitable corpora are usually of limited size. Moreover, for concrete domains such as biomedicine, corpora availability could be a problem due to the sensitivity of clinical data [26]. These problems hamper the accuracy and applicability of such corpora-based IC measures [43].

New *intrinsic* models have emerged to overcome the limitations of corpora-based IC calculus. They estimate the IC of a concept based on its taxonomic position in the ontology [41, 43, 52, 59]. These models consider that the taxonomic structure offered by an ontology is meaningfully organised according to the *cognitive saliency* principle [10]: concepts are taxonomically specialised when they must be differentiated from others. Then, assuming that the meaning of a concept is represented and bounded by its set of hyponyms, they compute the IC of a concept according to its number of specialisations. The rationale is that, as stated by Resnik in equation (5), general concepts with a large number of hyponyms will more probably (implicitly) occur in a corpus than more specific concepts with a restricted set of hyponyms.

Compared to corpora-based probability calculus, intrinsic models do not rely on tagged texts, so they provide a direct and fast solution to compute concept ICs. Moreover, empirical evaluations have underlined that intrinsic models are usually more accurate than corpora-based models for similarity assessments [41, 43, 52]. Several intrinsic IC (*iIC*) models that take advantage of these interesting properties have been proposed in recent years [52, 59]. Here we refer to the one proposed in [43]. This model incorporates additional semantic evidence in order to achieve better differentiation of concepts with the same number of hyponyms but different degrees of concreteness—this dimension is captured by their number of taxonomic ancestors or *subsumers*. Moreover, this is especially interesting in scenarios in which heterogeneous ontologies are used [42]. Indeed, it avoids depending on the inner-granularity of the taxonomic hierarchy by estimating $IC(c)$ from the subset of strict hyponyms of c that are leaves of the taxonomic tree,

denoted as *leaves*(*c*), rather than from the full set of hyponyms of *c*. More formally, it computes the *iIC*(*c*) by approximating its probability by the ratio between the size of *leaves*(*c*), as a measure of the generality of *c*, and the number of *subsumers* of *c* including itself (*subs*(*c*)), as a measure of the concreteness of *c*. This ratio is divided by *max_leaves*, which represent the number of leaves of the ontology, which acts as a normalisation factor to obtain values in the 0..1 range. Finally, 1 is added to the numerator and denominator to avoid *log*(*0*).

$$iIC(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subs(c)|}+1}{max\_leaves+1}\right) \quad (6)$$

Note that, as demonstrated in [43], *iIC*(.) is a monotonically decreasing function as one moves up in the taxonomy (i.e. ($\forall c_1 \in subs(c_2) \mid c_1 \neq c_2$)=> $iIC(c_1) < iIC(c_2)$).

3 Related works on semantic similarity from multiple ontologies

Substantial research has focused on the definition of semantic similarity measures, whereas much less has been devoted to tackling the similarity assessment problem in a multi-ontology context. Most of these latter works either propose methods to extend basic path-length measures [2, 50] or focus on feature-based measures [7, 34, 38], which usually consider ontological knowledge other than taxonomical (e.g. meronyms, concept glosses, related terms, etc.) that are much rarely available in ontologies [16].

Regarding IC-based measures, as stated in section 2.1, it is essential to discover suitable ancestors between different ontologies that can act as the MICA to enable cross-ontology similarity assessment. Two related works tackling this problem are further detailed below.

In [51], the authors consider that two ancestors are semantically equivalent only if they share the same textual label. The MICA of two concepts c_1 and c_2 is thus obtained by looking for a terminological match among their subsumers. Finally, the IC of this pair, representing the MICA of c_1 and c_2, is computed as the *iIC* of the concept of the pair with the highest number of hyponyms.

In [42], the authors distinguish three cases according to the ontology in which concepts c_1 and c_2 are found. Firstly, if both appear in a single ontology, the semantic similarity is computed, as detailed in section 2.1, using equation (6) to compute *iIC*. Secondly, if both appear at the same time in several ontologies, the authors retrieve the MICA corresponding to each of the ontologies and select the ontology containing the MICA of maximum IC to compute the semantic similarity, as done in the first case. Finally, if c_1 belongs to O_1 and c_2 belongs to O_2, the authors compare the labels of the subsumers of c_1 and c_2 and select, as their MICA, the most specific and terminologically matching subsumer pair $\langle c_i, c_j \rangle$ (i.e. c_i and c_j share a common label). The IC of the MICA is then computed as the maximum between the *iIC* of c_i and c_j estimated according to their respective ontologies.

The above approaches are hampered by the strict terminological matching applied to discover common subsumers. Note that this remark also holds for works focusing on other similarity paradigms where common ancestors between ontologies are also discovered based on concept label matching. This was the case of approaches dealing with edge-counting measures [2, 7] or feature-based similarities [34, 38] in a multi-ontology setting.

In fact, the accuracy of terminological matching between concept labels is seriously limited by the fact that ontologies rarely model knowledge in the same way, since knowledge engineers create ontologies according to their specific needs, viewpoints and application domains. This is

especially noticeable for the most abstract concepts of the taxonomy, i.e. those from which the MICA is usually retrieved. Unlike concrete concepts that, due to their specificity, are commonly referred with unambiguous textual labels (e.g. concrete disease names), general concepts usually refer to ad-hoc abstractions (e.g. *entity*, *concept*, *substance*, *procedure*, etc.), whose labels are hard to match due to their non-consensual nature [50]. Moreover, the level of detail and granularity of the inner taxonomic structure usually vary even among ontologies modelling the same knowledge domain. In a multi-ontology context, methods thus need to be able to deal with significantly different taxonomic structures whose concepts do not match in an ideal 1:1 mapping. Since concept subsumers do not perfectly overlap, their textual labels are unlikely to match. For example, in Figure 1, there are clear differences in the granularities of the taxonomical hierarchies of MeSH and SNOMED-CT associated with the *Heart* concept.

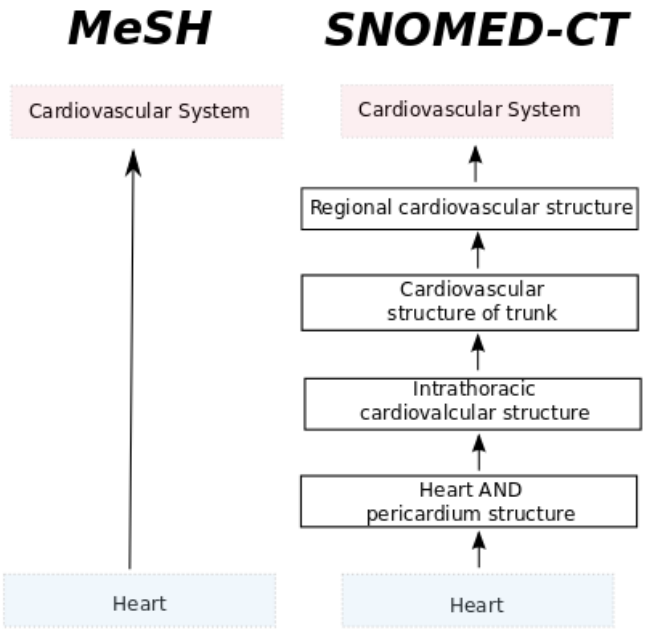

Figure 1. Excerpt of MeSH and SNOMED-CT taxonomies corresponding to the concept *Heart*.

The above issues hamper the accuracy of the straightforward terminological matching of concept subsumers employed by related works, hence limiting the precision of their subsequent similarity assessments. Indeed, for such approaches, if no terminological match is found

between subsumers, the compared concepts will be assessed as maximally distant, since no evidence of commonalty between the two ontologies has been found.

Note that, even if a pair of terminologically equivalent subsumers is found, the similarity estimation may be underestimated if there is a more specific pair of concepts, which are highly equivalent despite having different labels (e.g., *cancer*/*neoplasm*). For example, in Figure 2, a terminological matching strategy will clearly only be able to select the pair of subsumers labelled as *disease* as the MICA between *Intracranial Hemorrhage* (in SNOMED-CT) and *Brain neoplasms* (in MeSH). However, the pair of subsumers *Disease of head* (SNOMED-CT) and *Brain Diseases* (MeSH) would serve as better evidence of commonality for similarity assessment. In this case, since more abstract subsumers with lower ICs are considered, similarity is underestimated.

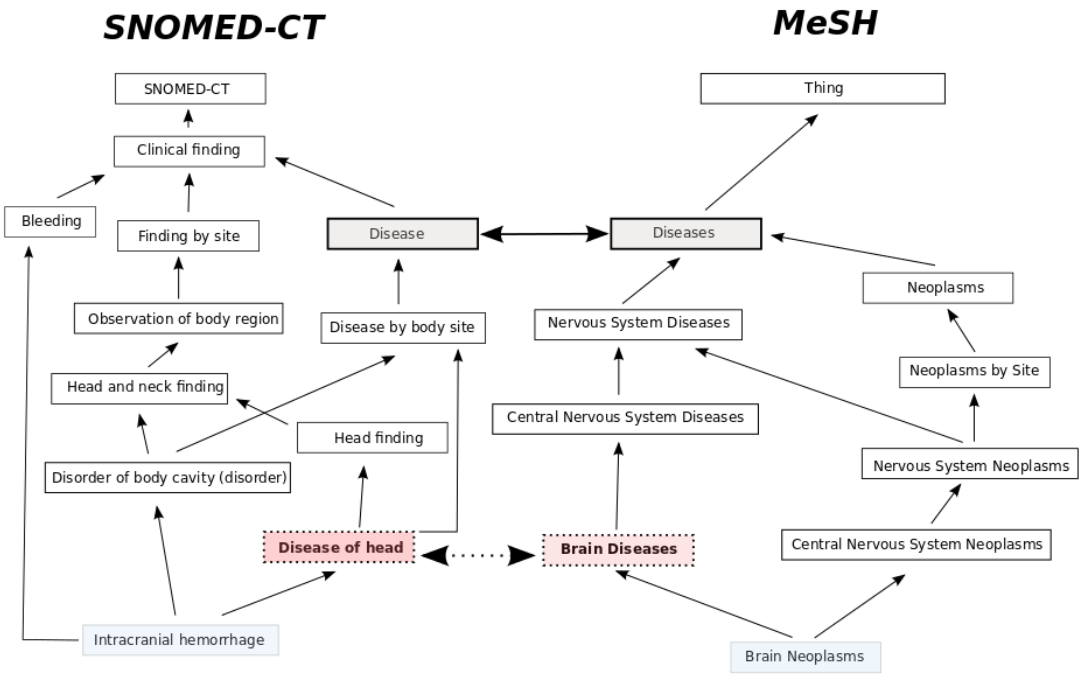

Figure 2. Taxonomic structures for *Intracranial Hemorrhage* (in SNOMED-CT) and *Brain neoplasms* (in MeSH).

The above examples highlight the need for a method able to detect, beyond terminological matchings, equivalences between suitable pairs of subsumers that could improve semantic similarity assessments in a multi-ontology scenario.

4 A method to improve IC-based similarity assessment from multiple ontologies

In this section, we present a method to enable accurate IC-based similarity calculus when concepts belong to different ontologies. Our method goes beyond the terminological matching used in related works and is able to discover semantically similar (but not necessarily terminologically identical) subsumers between different ontologies. To do so, and in line with the notion of IC-based similarity, our approach relies on the foundations of Information Theory and the notion of *mutual information* to quantify the degree of semantic equivalence between subsumer pairs from different ontologies. From these, the *most equivalent* pair is taken as MICA. Then, its IC is determined and used to estimate the semantic similarity between the two concepts using standard measures (e.g. equations 2, 3 or 4).

4.1 Basic Definitions

Definition 1. Let $C(O)$ be the set of concepts of an ontology O. The *subsumption relationship*, denoted as $\leq$ hereafter, defines a partial order on the concepts $C(O)$. Given two concepts c_i and c_j from O, $c_i \leq c_j$ accounts for the fact that c_j is an ancestor of c_i in O or $c_i=c_j$.

Definition 2. The set of *subsumers* of a concept c in the ontology O is defined as:

$$subs_O(c) = \{s \in O \mid c \leq s\}$$

Definition 3. The set of *hyponyms* of a concept *c* in the ontology *O* can be similarly defined as:

$$hypo_O(c) = \{h \in O \mid h \leq c\}$$

Note that, according to these definitions, a concept is included both in its sets of *subsumers* and *hyponyms*.

4.2 Discovering the MICA across different ontologies

To discover an appropriate MICA for a pair of concepts c_1 and c_2, where $c_1 \in O_1$ and $c_2 \in O_2$, their sets of subsumers (i.e. $subs_{O1}(c_1)$ and $subs_{O2}(c_2)$) should be evaluated in order to find the most semantically equivalent pair. To do so, we first search for the most informative pair of subsumers from $subs_{O1}(c_1)$ and $subs_{O2}(c_2)$ with identical labels, which we will use as a boundary during MICA search. The search space is not bounded if no terminological matchings are found. Formally:

Definition 4. Given $c_1 \in O_1$ and $c_2 \in O_2$, their most informative pair of terminologically-matched subsumers, denoted here as $\langle ms_1, ms_2 \rangle$, are defined as:

$$\langle ms_1, ms_2 \rangle = \begin{cases} \underset{\forall \langle s_i, s_j \rangle \mid \langle s_i \in subs_{O_1}(c_1), s_j \in subs_{O_2}(c_2) \rangle}{\arg\max} (iIC(s_i) + iIC(s_j)) & , if\ \exists s_i, s_j \mid s_i \equiv s_j \\ \langle root\_node(O_1), root\_node(O_2) \rangle & , otherwise \end{cases}$$

where the '≡' operator indicates that s_i and s_j at least share an identical textual label, while considering synonym sets associated with each concept, if available. The maximisation of the sum of intrinsic ICs of matched subsumers ensures that, in case of multiple terminological matches, the most informative pair is taken.

Definition 5. The set of subsumer pairs that are candidates to be the MICA of c_1 and c_2 (contained in the set named *cMICA*) are the tuples $\langle cs_i, cs_j \rangle$ resulting from the Cartesian product

of the set of subsumers in $subs_{O1}(c_1)$ and in $subs_{O2}(c_2)$, which are taxonomically equal or below $\langle ms_1, ms_2 \rangle$:

$$cMICA(c_1, c_2) = \left\{ \langle cs_i, cs_j \rangle \in \left\{ subs_{O_1}(c_1) \times subs_{O_2}(c_2) \right\} | \left(cs_i \leq ms_i \right) \wedge \left(cs_j \leq ms_j \right) \right\}$$

Note that, according to the subsumption relation $\leq$ (Definition 1), pairs $\langle c_1, c_2 \rangle$ and $\langle ms_1, ms_2 \rangle$ are included in $cMICA(c_1,c_2)$ and are thus considered as potential MICA for c_1, c_2.

Among the candidate pairs included in $cMICA(c_1, c_2)$, we select as $MICA(c_1, c_2)$ the pair with the highest degree of *semantic equivalence*, i.e. the pair $\langle cs_i, cs_j \rangle$ such that cs_i subsumes *most of the semantics* of cs_j and vice-versa. The rationale for this criterion is that ideal common subsumers are those that can generalise the *same* semantics so that they are interchangeable with regard to the notion of concept subsumption. However, we take the pair with the *maximum semantic equivalence* as the best available MICA since a perfect semantic equivalence is rare when dealing with heterogeneous knowledge sources.

By relying on the foundations of Information Theory, we use the *Mutual Information* (MI) of subsumer pairs as a proxy of their equivalence. The instantiation of MI for each subsumer pair $\langle cs_i, cs_j \rangle$ results in the *Point-wise Mutual Information* (PMI) [14]. Similar to the IC calculus, the PMI between two concepts can be computed according to their probabilities of occurrence and co-occurrence. Formally, the PMI quantifies the difference between the probability of concept co-occurrence given their joint and marginal probabilities [14].

$$PMI(cs_i; cs_j) = \log \frac{p(cs_i, cs_j)}{p(cs_i)\, p(cs_j)} \tag{7}$$

Given the above expression, PMI=0 means that the two subsumers are completely independent, whereas increasing positive values indicate an increasing degree of association or semantic

mutuality. On the contrary, negative values reflect mutual exclusion, which is quite uncommon among concepts or words, since most of them tend to be semantically correlated up to a certain degree [3]. Also note that the PMI function is symmetric.

Since $p(cs_i,cs_j)$ can be rewritten in terms of conditional probability as $p(cs_i|cs_j)p(cs_j)$, it can be deduced from equation (7) that:

$$PMI(cs_i;cs_j) = \log\frac{p(cs_i \mid cs_j)\,p(cs_j)}{p(cs_i)\,p(cs_j)} = \log\frac{p(cs_i \mid cs_j)}{p(cs_i)}$$

Considering that $IC(c)=-log(p(c))$, by extending this notation, we denote $IC(cs_i|\ cs_j)=-log(p(cs_i|\ cs_j))$ the *conditional information content* provided by cs_i considering that we know cs_j. By using this notation, PMI can be rewritten as:

$$PMI(cs_i;cs_j) = \log\frac{p(cs_i \mid cs_j)}{p(cs_i)} = \log(\,p(cs_i \mid cs_j)) - \log(\,p(cs_i)) = IC(cs_i) - IC(cs_i \mid cs_j)$$

Since $p(cs_i,cs_j)$ is also equal to $p(cs_j|cs_i)p(cs_i)$, a similar transformation highlights the symmetry of the PMI measure and its relationships with the IC:

$$PMI(cs_i;cs_j) = IC(cs_i) - IC(cs_i \mid cs_j) = IC(cs_j) - IC(cs_j \mid cs_i)$$

This result emphasises the fact that $PMI(cs_i;cs_j)$ quantifies how much information cs_j tells us about cs_i and vice-versa. Hence, a maximal PMI value states that both concepts provide *all* the information from each other.

A common criticism concerning PMI is that it tends to provide relatively high scores for rare events [12]. For example, we have $p(cs_i)=p(cs_j)=p(cs_i,cs_j)$ when two terms only occur together and it then follows from equation (7) that $PMI(cs_i,cs_j)=-\log(p(cs_i,\ cs_j))$. This means that, for perfectly correlated concepts, their PMI value will be higher when they appear less frequently.

Moreover, PMI has no *fixed* upper (respectively lower) bounds, which complicates its interpretation since it is thus hard to know from a given PMI value if two concepts are almost perfectly associated (respectively almost independent). These problems may be partly solved by using the *Normalised Pointwise Mutual Information* (NPMI). Indeed, NPMI values are bounded within the interval [-1..1] and are less impacted by low frequency data [12]. NPMI normalisation is done by dividing the PMI ratio by the actual probability of co-occurrence between the two terms:

$$NPMI(cs_i;cs_j) = \frac{PMI(cs_i,cs_j)}{-\log p(cs_i,cs_j)} = \frac{\log \frac{p(cs_i,cs_j)}{p(cs_i)p(cs_j)}}{-\log p(cs_i,cs_j)} \quad (8)$$

NPMI results in a maximum value of 1 for perfect correlation, a minimal value of -1 for mutually exclusive concepts, and a value of 0 for independent concepts since their PMI is null.

Given the above arguments and properties, NPMI provides a sound way to measure concept mutuality. In the next section, we detail the precise instantiation of this NPMI measure to assess our pursued degree of subsumer equivalence in a multi-ontology context.

4.3 Probability calculus for measuring semantic equivalence

The choice of the estimation used to compute the co-occurrence probability $p(cs_i,cs_j)$ is crucial to ensure that $NPMI(cs_i,cs_j)$ correctly reflects the degree of semantic equivalence between two concepts. As for the IC estimation, this probability can be estimated using suitable corpora, by counting the number of simultaneous appearances of those two concepts. In our setting, this results in two main issues:

- First, as already discussed in section 2.2, corpora-based probability calculus is hampered by data sparseness and restricted by corpora availability.

- Secondly, since term co-occurrences are not usually disambiguated (i.e. the kind of relationship, taxonomic or non-taxonomic, underlying the co-occurrence is not defined), the co-occurrence frequency gives us a measure of semantic proximity, i.e. a general degree of *association* or *relatedness* between concepts [11], which is a more general notion than the *semantic equivalence* that we pursue. Indeed, co-occurrence frequencies and PMI/NPMI measures based on them have been applied to evaluate heterogeneous types of semantic association such as word collocation [12, 48], taxonomic subsumption [56], and a variety of non-taxonomic relationships [39, 49]. In the context of our problem, the use of co-occurrence frequency to estimate $p(cs_i,cs_j)$ may result in high NMPI scores for related but not equivalent subsumer pairs. This issue becomes problematic since we are interested in unravelling equivalent taxonomic ancestors to find a suitable MICA. For example, let us consider two possible concepts: *cancer* and *chemotherapy*. Assuming the availability of an appropriate corpus, their degree of textual co-occurrence is likely to be high, resulting in an also high NPMI value. However, since *chemotherapy* is a common treatment for *cancer*, those numerous co-occurrences reflect a close relationship rather than a degree of semantic equivalence.

We propose to tackle those problems by using probability estimations derived from the taxonomy in which concepts are modelled. The probabilities of individual concepts are computed intrinsically, according to the premises of the intrinsic IC calculus discussed in section 2.2. Thus, we propose an intrinsic approximation of the term co-occurrence probability that *solely* evaluates ontological evidence of *semantic equivalence*. Specifically, as in intrinsic IC models, we rely on the fact that the meaning of a subsumer is defined and bounded by its set of hyponyms [43, 52]. For instance, the meaning of *body part* is the result of the aggregation of all its specialisations (i.e. anatomical entities). Hence, the number of *shared hyponyms* between two subsumers gives us a good clue to their semantic equivalence, i.e. as the overlap between their hyponym sets increases, their meanings become more equivalent. Note that this notion of

equivalence, which is the core of our subsumer matching method, differs from that of *similarity* quantified by IC-based measures, since two sibling concepts (e.g. *breast cancer* and *lung cancer*) may be highly similar (according to their IC-based similarity) while sharing no hyponyms and, hence, being completely disjoint. In fact, hyponym overlapping would only occur between equivalent concepts or strict taxonomic ancestors or specialisations, which are the cases upon which our subsumer matching method focuses.

Formally, we approximate the joint probability of two concepts $cs_i \in O_1$ and $cs_j \in O_2$ by the number of their shared hyponyms divided by the number of distinct concepts present in O_1 and O_2 (i.e. $C(O_1)$ and $C(O_2)$, respectively), so that subsumer pairs from ontologies of different sizes can be fairly compared.

$$p(cs_i, cs_j) \cong \frac{|\, hypo_{O_1}(cs_i) \cap hypo_{O_2}(cs_j) \,|}{|\, C(O_1) \cup C(O_2) \,|} \tag{9}$$

Since we are considering intersection and union of concept sets coming from distinct ontologies, we need to extend those operators for this multi-ontology context. According to basic set theory while evaluating shared concepts based on terminological matchings, denoted as '≡', we define the size of the intersection between a set S_1 of concepts of O_1 and a set S_2 of concepts of O_2 as:

$$|\, S_1 \cap S_2 \,| = \frac{|\, \{ c_x \in S_1 \mid \exists c_y \in S_2 \wedge (c_x \equiv c_y) \} \,| + |\, \{ c_y \in S_2 \mid \exists c_x \in S_1 \wedge (c_y \equiv c_x) \} \,|}{2}$$

The size of the union is then simply defined as the complement of the intersection:

$$|\, S_1 \cup S_2 \,| = |\, S_1 \,| + |\, S_2 \,| - |\, S_1 \cap S_2 \,|$$

Note that the thus defined union and intersection sizes are not necessarily integer values but ensure consistent results even with ontologies that have polysemic concept labels and

heterogeneous granularities for which 1:M or even N:M matchings between concepts can be found. For instance, if $S_1=\{c_a,c_b\}$; $S_2=\{c_e\}$ and $c_a \equiv c_e$ and $c_b \equiv c_e$, then $|S_1 \cap S_2|=1.5$ and $|S_1 \cup S_2|=1.5$. Note also that, thanks to the higher cardinality of hyponym sets and the less ambiguity inherent to their higher degree of concreteness, the chance of obtaining a representative number of matchings increases in comparison with subsumer sets.

In accordance with equation (9), we define the marginal probability of an individual subsumer cs_x as:

$$p(cs_x) \cong \frac{|\,hypo_{O_y}(cs_x)\,|}{|\,C(O_1) \cup C(O_2)\,|} \tag{10}$$

Given the above instantiations of joint and marginal probabilities, we define the intrinsic NPMI of a subsumer pair as follows.

Definition 6. Given $cs_i \in O_1$ and $cs_j \in O_2$, their *intrinsic NPMI (or iNPMI)* is defined as:

$$iNPMI(cs_i;cs_j) = \frac{\log \frac{p(cs_i,cs_j)}{p(cs_i)\,p(cs_j)}}{-\log p(cs_i,cs_j)} = \frac{\log \frac{\left(\frac{|hypo_{O_1}(cs_i) \cap hypo_{O_2}(cs_j)|}{|C(O_1) \cup C(O_2)|}\right)}{\left(\frac{|hypo_{O_1}(cs_i)|}{|C(O_1) \cup C(O_2)|}\right) \times \left(\frac{|hypo_{O_2}(cs_j)|}{|C(O_1) \cup C(O_2)|}\right)}}{-\log\left(\frac{|hypo_{O_1}(cs_i) \cap hypo_{O_2}(cs_j)|}{|C(O_1) \cup C(O_2)|}\right)} \tag{11}$$

Numerically, an $iNPMI(cs_i;cs_j)=0$ indicates that subsumers cs_i and cs_j have no overlapping and, therefore, that these two concepts cannot serve as MICA. On the contrary, an $iNPMI(cs_i;cs_j)$ value close to 1 indicates that cs_i and cs_j are close to being semantically equivalent since they share most of their hyponyms. In our approach, the MICA pair selected for two concepts is thus the candidate with the highest *iNPMI* value.

Definition 7. The MICA for $c_1 \in O_1$ and $c_2 \in O_2$ is a pair of subsumers from *cMICA* that fulfils:

$$M = \underset{\forall \langle cs_i, cs_j \rangle \in cMICA}{\arg\max} \{iNPMI(cs_i, cs_j)\}$$

Since several pairs (in *M*) may maximise the expression, as the final MICA, we take the pair whose sum of *iIC* values is maximum (i.e. the most informative one, in coherency with the notion of MICA).

$$MICA(c_1, c_2) = \underset{\forall \langle cs_i, cs_j \rangle \in M}{\arg\max} \{iIC(cs_i) + iIC(cs_j)\}$$

The above-described method can be generalised if c_1 and/or c_2 belong to several disjoint ontology sets (e.g. c_1 belongs to O_1 and O_3, and c_2 belongs to O_2 and O_4). In that case, the proposed method is applied 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) and the MICA is selected as the subsumer pair belonging to the ontology pair that produces the highest *iNPMI* value. The rationale is that as the *iNPMI* increases, the subsumers become more semantically equivalent and hence the selected MICA is more suitable for the compared concepts. Note that the normalised and bounded outputs provided by the NPMI measure (as discussed in section 4.2) are quite convenient to fairly compare NPMI values computed from different ontology pairs, regardless of their sizes and granularity degrees.

4.4 IC-based similarity calculus

The result of the above-proposed $MICA(c_1,c_2)$ identification procedure is a pair $\langle cs_i, cs_j \rangle$, where $cs_i \in O_1$ and $cs_j \in O_2$. However, the IC for the $MICA(c_1,c_2)$ should be uniquely defined to be useful to assess the semantic similarity between c_1 and c_2 using classical IC-based measures (see section 2.1).

Moreover, as stated in section 2.2, the IC value of this MICA should be lower than any of its hyponyms in order to ensure the consistency of IC-based similarity measures [37]. To ensure

that this property will be fulfilled in our multi-ontology setting, we define the intrinsic IC of the selected MICA as the minimum *iIC* value (computed as in equation (6)) of each concept in $\langle cs_i, cs_j \rangle$ computed from its ontology.

Definition 8. Given the $MICA(c_1,c_2)=\langle cs_i,cs_j \rangle$ with $cs_i \in O_1$ and $cs_j \in O_2$, its *iIC* is defined as:

$$iIC(MICA(c_1, c_2)) = \min(iIC(cs_i), iIC(cs_j)) \quad (12)$$

Proposition 1. Definition 8 fulfils that $iIC(MICA(c_1,c_2))<iIC(c_1)$ and $iIC(MICA(c_1,c_2))<iIC(c_2)$.

Proof. The proof is trivial, considering that: i) $MICA(c_1,c_2)= \langle cs_i,cs_j \rangle$, ii) $c_1 \leq cs_i$ and $c_2 \leq cs_j$, iii) *iIC* calculus detailed in equation (6) already fulfils that $iIC(cs_i)<iIC(c_1)$ and $iIC(cs_j)<iIC(c_2)$. Hence, according to definition 8: $iIC(MICA(c_1,c_2))=min(iIC(s_i),iIC(s_j))$, we have: $iIC(MICA(c_1,c_2)) \leq iIC(s_i)<iIC(c_1)$ and $iIC(MICA(c_1,c_2)) \leq iIC(s_j)<iIC(c_2)$. □

5 Evaluation

In this section, we evaluate the proposed method in comparison with related works. Since our final goal is to enable a precise IC-based assessment of similarity in a multi-ontology setting, we focused on cases where each concept to evaluate belongs to a different ontology. In such multi-ontology scenarios, similarity assessments directly depend on the adequacy of the subsumer pair selected as MICA and the subsequent IC calculus, as detailed in section 4. Hence, by quantifying the accuracy of the similarity assessment, we also indirectly test the relevance of our MICA identification strategy. The accuracy of the proposed method is compared with those of related works also focusing on multi-ontology IC-based similarity calculus [42, 51], and with results obtained in an "ideal" mono-ontology setting (i.e. when similarities are computed from a single ontology).

The most widely used way to objectively evaluate the accuracy of semantic similarity assessments (and, thus, of strategies to enable them in a multi-ontology setting) consists on comparing them with human judgements of similarity. More specifically, the Pearson's correlation is usually computed to measure the agreement between human ratings and computerised assessments for a benchmark composed of pairs of terms [2, 11, 35, 40, 44]. A correlation value near 1 indicates that both ratings are very close and, hence, that the score produced by the evaluated measure accurately reflects human judgements of similarity. The idea is that, since the goal of any application relying on similarity assessments (e.g. semantic disambiguation, document classification, etc.) is to mimic human decisions, similarity measures should capture and reproduce human similarity ratings as much as possible [31]. In fact, several empirical works have already shown the close relationship between the accuracy of similarity measures (in terms of their correlation with human ratings) and the performance of specific applications that rely on them. For example, in [8], the authors evaluated several similarity measures that were applied to data clustering/classification; results showed that best classifications were obtained by the measures that best correlated with human ratings. More specifically, in [5, 9], the authors showed in the context of data clustering that the multi-ontology similarity measures that best correlated with human ratings also enabled most accurate classifications. In a recent study [28], different similarity paradigms were evaluated within the context of semantic disambiguation, reaching the same conclusions. Finally, in [13] a similar comparative study was performed applying similarity measures to spelling correction. Because of the above arguments, the calculus of measures' correlation against human ratings for well-known benchmarks [31, 33] of term pairs have become a *de facto* evaluation protocol within the semantic similarity research community. This is because it facilitates the reproducibility of the results and enables a direct comparison of different methods. Moreover, such an evaluation protocol gives a clear insight on the expected performance of similarity methods in specific

applications relying on them but without being tied to a concrete task. Thus, our evaluation will be performed in the same way.

Evaluated methods propose different solutions to identify the MICA of concepts from different ontologies and to estimate its IC, which can then be used to compute the semantic similarity using IC-based measures like those introduced in section 2.1. In our tests, the IC of individual concepts was computed intrinsically according to equation (6). Moreover, since the relative accuracies of tested methods may depend on the IC-based measure chosen for similarity calculus, we tested them with several measures: Resnik's [37], Lin's [25] and Jiang and Conrath's [22].

The evaluation was conducted on biomedical datasets because of the availability of different ontologies and similarity benchmarks in this field. In particular, *Systematized Nomenclature of Medicine, Clinical Terms* (SNOMED-CT) and the *Medical Subject Headings* (MeSH) knowledge sources have been used as ontologies: SNOMED-CT [53] covers more than 360,000 medical concepts classified in 18 partially overlapping hierarchies, whereas the MeSH [29] provides around 25,000 medical and biological concepts hierarchically classified in 16 categories. Methods have been compared by computing similarity correlations using two biomedical similarity benchmarks, i.e. the one proposed by Pedersen *et al.* [33] and the one by Pakhomov *et al.* [31]. The former consists of a set of medical term pairs whose similarity was assessed by a group of medical experts from the Mayo Clinic: 9 medical coders who were introduced to the notion of semantic similarity and 3 physicians who rated terms without any special training. Term pairs included in the benchmark were specifically selected by the authors to maximise the inter-rating agreement, resulting in a correlation value of 0.68 obtained for physicians and of 0.78 for coders. The benchmark built by Pakhomov *et al.* consists of a set of concept pairs associated with similarity and relatedness ratings given by four medical residents from the University of Minnesota. We took the similarity ratings since we were focusing

specifically on semantic similarity. Note that for the Pakhomov *et al*. benchmark, the inter-rating agreement (0.47) is significantly lower than that obtained for the dataset of Pedersen *et al*.

Even though such benchmarks are intended to evaluate similarity measures in a mono-ontology setting, related works have already used them in a multi-ontology framework by artificially considering that *each* term of each pair belongs to a *different* ontology. According to the same protocol as in [7, 42], we took the 25 term pairs from the Pedersen *et al*. benchmark and the 150 concept pairs from the Pakhomov *et al*. benchmark, such that both elements of the pair could be found in SNOMED-CT as well as in MeSH. Hence, it was possible, for the 175 pairs considered in our evaluation procedure, to assess their pairwise similarity in a mono-ontology setting. Since similarity estimations are obviously harder and more precarious in a multi-ontology setting, mono-ontology results give us approximated upper bounds of the *best* accuracies we can expect in a multi-ontology setting. For each benchmark, we conducted two different multi-ontology evaluations. In the first case, referred as "SNOMED-CT + MeSH", the first concept of each pair was retrieved from SNOMED-CT and the second one from MeSH. Whereas in the second case, referred to as "MeSH + SNOMED-CT", the first concept was retrieved from MeSH and the second one from SNOMED-CT.

The SNOMED-CT release of July 2012 (20120731) and the MeSH 2013 release were used for the evaluation. Semantic measures were implemented using the open source Semantic Measures Library[i] [20]. Evaluation details, source code and associated datasets can be downloaded from the dedicated webpage[ii].

Tables 2 and 3 show the correlation values obtained by each IC-based similarity measure for the two benchmarks. Tables show the cases in which (i) both concepts are retrieved from

[i] http://www.semantic-measures-library.org
[ii] http://www.lgi2p.ema.fr:8090/~sharispe/publications/IS2013/

SNOMED-CT, (ii) both are evaluated in MeSH, and (iii) when each concept is considered in a different ontology (SNOMED-CT + MeSH and MeSH + SNOMED-CT), using the MICA discovery and calculus strategies of Sanchez and Batet [42], Saruladha *et al.* [51] and the one presented in this paper.

Table 2. Correlation values of different IC-based measures against human ratings for term pairs extracted from the Pedersen *et al.* benchmark in mono and multi-ontology scenarios. Rows in **boldface** show the results of our proposal.

IC-based measure	*Ontologies*	*MICA discovery*	*Physicians*	*Coders*	*Both*
Resnik	SNOMED-CT	None	0.553	0.598	0.602
	MeSH	None	0.608	0.668	0.670
	SNOMED-CT + MeSH	Sánchez and Batet	0.489	0.544	0.542
	SNOMED-CT + MeSH	Saruladha et al.	0.474	0.546	0.535
	SNOMED-CT + MeSH	**This work**	**0.617**	**0.624**	**0.649**
	MeSH + SNOMED-CT	Sánchez and Batet	0.444	0.534	0.512
	MeSH + SNOMED-CT	Saruladha et al.	0.432	0.536	0.508
	MeSH + SNOMED-CT	**This work**	**0.562**	**0.639**	**0.632**
Lin	SNOMED-CT	None	0.566	0.628	0.625
	MeSH	None	0.614	0.674	0.676
	SNOMED-CT + MeSH	Sánchez and Batet	0.512	0.561	0.561
	SNOMED-CT + MeSH	Saruladha et al.	0.501	0.569	0.560
	SNOMED-CT + MeSH	**This work**	**0.637**	**0.654**	**0.674**
	MeSH + SNOMED-CT	Sánchez and Batet	0.446	0.542	0.517
	MeSH + SNOMED-CT	Saruladha et al.	0.432	0.543	0.511
	MeSH + SNOMED-CT	**This work**	**0.561**	**0.648**	**0.637**
Jiang and Conrath	SNOMED-CT	None	0.538	0.612	0.602
	MeSH	None	0.618	0.670	0.676
	SNOMED-CT + MeSH	Sánchez and Batet	0.514	0.573	0.569
	SNOMED-CT + MeSH	Saruladha et al.	0.505	0.580	0.569
	SNOMED-CT + MeSH	**This work**	**0.637**	**0.651**	**0.673**
	MeSH + SNOMED-CT	Sánchez and Batet	0.423	0.527	0.498
	MeSH + SNOMED-CT	Saruladha et al.	0.404	0.524	0.487
	MeSH + SNOMED-CT	**This work**	**0.542**	**0.638**	**0.622**

Table 3. Correlation values of different IC-based measures against human ratings for term pairs extracted from the Pakhomov *et al*. benchmark in mono and multi-ontology scenarios. Rows in **boldface** show the results of our proposal.

IC-based measure	*Ontologies*	*MICA discovery*	*Experts*
Resnik	SNOMED-CT	None	0.513
	MeSH	None	0.511
	SNOMED-CT + MeSH	Sánchez and Batet	0.315
	SNOMED-CT + MeSH	Saruladha et al.	0.305
	SNOMED-CT + MeSH	**This work**	**0.493**
	MeSH + SNOMED-CT	Sánchez and Batet	0.260
	MeSH + SNOMED-CT	Saruladha et al.	0.243
	MeSH + SNOMED-CT	**This work**	**0.429**
Lin	SNOMED-CT	None	0.505
	MeSH	None	0.519
	SNOMED-CT + MeSH	Sánchez and Batet	0.320
	SNOMED-CT + MeSH	Saruladha et al.	0.310
	SNOMED-CT + MeSH	**This work**	**0.505**
	MeSH + SNOMED-CT	Sánchez and Batet	0.257
	MeSH + SNOMED-CT	Saruladha et al.	0.244
	MeSH + SNOMED-CT	**This work**	**0.447**
Jiang and Conrath	SNOMED-CT	None	0.456
	MeSH	None	0.520
	SNOMED-CT + MeSH	Sánchez and Batet	0.313
	SNOMED-CT + MeSH	Saruladha et al.	0.232
	SNOMED-CT + MeSH	**This work**	**0.505**
	MeSH + SNOMED-CT	Sánchez and Batet	0.257
	MeSH + SNOMED-CT	Saruladha et al.	0.199
	MeSH + SNOMED-CT	**This work**	**0.448**

First, to demonstrate the statistical significance of the reported results, we computed the *p-value* of the obtained correlations. Since the *p-value* of a statistic (in this case the correlation) measures the probability that observations occurred by chance, a low value is desirable. In all cases, *p-values* of reported correlations were lower than 0.01, which states that reported results are statistically significant since the probability of a random chance is less than 1%.

Analysing the results, we observe that methods based only on terminological matchings resulted in correlation values that were below the worst mono-ontology setting, i.e. those of SNOMED-CT in these tests. Given that, in our testing protocol, all methods relied on the same IC calculus (equation (6)), the differences between the method of Sanchez and Batet and that of Saruladha *et al.* were very minor. In fact, since both methods look for the most specific pair of terminologically matching subsumers, the only practical difference for the evaluated scenarios regards the criterion of the IC calculus for the discovered pair. Indeed, Saruladha *et al*. select the minimum IC from the matched pair, whereas Sanchez and Batet take the maximum value. The difference in performance between those two methods and the mono-ontology settings strongly depends on the considered dataset and IC-based measures. In some cases, this difference could be small (e.g. 0.534-0.546 vs. 0.598-0.668 for Resnik's measure and the Pedersen *et al.* coder ratings) or significantly large (e.g. 0.199-0.313 vs. 0.456-0.520 for Jiang and Conrath's measure and the Pakhomov *et al.* expert ratings, which represents a more challenging dataset). As discussed in section 3, those two methods are hampered by the fact that, in many cases, the terminologically matching subsumer pair is more abstract than it should be, and this results in an underestimation of the true similarity between the compared concepts. This issue is specifically tackled by our approach, which looks for a pair of subsumers with a higher degree of semantic equivalence than the terminologically matched one. A more suitable MICA pair can therefore be discovered, hence improving similarity assessments (see the example in Figure 2).

Indeed, correlations obtained using our method noticeably improve those of related works (e.g. 0.624-0.639 vs. 0.534-0.546 for Resnik's measure and the Pedersen *et al.* coder ratings, and 0.448-0.505 vs. 0.199-0.313 for Jiang and Conrath's measure and the Pakhomov *et al*. expert ratings). In fact, these results are close to those obtained in mono-ontology contexts, e.g. 0.624-0.639 vs. 0.598-0.668 for Resnik's measure and the Pedersen *et al.* coder ratings, and 0.448-0.505 vs. 0.456-0.520 for Jiang and Conrath's measure and the Pakhomov *et al*. expert ratings. These differences are also more uniform for all measures and datasets than those of related

works. Recall that, as stated above, correlation values reported in both tables for mono-ontology settings give us a reasonable approximation of the best correlation that can be achieved in multi-ontology scenarios. These results suggest that our strategy is close to optimal and leave little room for improvement.

Regarding IC-based similarity, our results confirm that Lin's and Jiang and Conrath's measures tend to lead to better results than those of Resnik. This is expected as the Resnik measure, unlike the two other measures, associates the same similarity to any concept pair with identical MICA, regardless of the IC of the compared concepts.

Note, finally, that computed similarities are more congruent with human ratings for the Pedersen *et al*. benchmark than for that of Pakhomov *et al.*. This result is coherent with the difference in inter-human agreement figures for the two benchmarks: 0.68-0.78 for Pedersen *et al*. vs. 0.47 for Pakhomov *et al*. The influence of the reliability of human ratings is also evident with the Pedersen *et al.* benchmark, where computed similarities are better correlated with the coders' ratings (which are more consistent, i.e. inter-rating agreement of 0.78) than with the physicians' rating (which are less consistent, i.e. inter-rating agreement of 0.68). The higher inter-rating agreement among coders is certainly related to the fact that they were trained on the notion of semantic similarity, whereas the physicians rated term pairs without previous training [33].

6 Conclusions

Due to the growing importance and availability of textual data in the so-called Information Society, semantic similarity, which represents the backbone of text understanding in many contexts, has become a hot topic. Many ontology-based similarity measures have been developed in recent years due to the availability of large and well-established ontologies. From

these, measures based on intrinsic IC computation models have achieved state-of-the art results [17, 40, 41], also providing a direct and unconstrained way to compare textual entities. However, the applicability of these measures is hampered by the fact that they were designed to deal with a *single* ontology. This constitutes a serious limitation given the increasing importance of distributed and multi-disciplinary scenarios in which multiple heterogeneous knowledge bases are available [1, 5, 23].

We propose a method to enable accurate IC-based similarity assessments from multiple ontologies, thus overcoming the above shortcoming. Our approach, grounded on the foundations of the Information Theory, looks for the available subsumer pair that can act as the best MICA for the compared concepts. Contrary to related works based solely on the lexical resemblance of subsumer labels, our method proposes a way to measure their degree of equivalence from a semantic standpoint. As a result, we discover subsumer pairs that better represent the commonalities of the compared concepts and that enable more accurate similarity assessments. The empirical evaluation, carried out on several well-established benchmarks, ontologies and measures, sustained the theoretical hypothesis: our method achieved similarity results that correlated significantly better with human ratings than those of related works, and that were very close to those obtained in the "optimal" mono-ontology setting.

As future work, we plan to evaluate our method in other domains and ontologies. We also plan to apply it to specific tasks dealing with multiple knowledge sources such as the classification of heterogeneous datasets [4, 9].

Acknowledgements

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

C03-01, ICWT TIN2012-32757, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004 and BallotNext IPT-2012-0603-430000), by the Government of Catalonia (under grant 2009 SGR 1135), by the AvieSan national program (French national alliance for life sciences and health) and by the French Agence Nationale de la Recherche 'Investissement d'avenir/Bioinformatique' (ANR-10-BINF-01-02 'Ancestrome') .

References

[1] M. Aaltonen, J. Holmström, Multi-ontology topology of the strategic landscape in three practical cases, Technological Forecasting and Social Change, 77 (2010) 1519–1526.

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

[3] B. Anandan, C. Clifton, Significance of Term Relationships on Anonymization, in: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03, IEEE Computer Society, Washington, DC, USA, 2011, pp. 253-256.

[4] M. Batet, Ontology based semantic clustering, AI Communications, 24 (2011) 291-292.

[5] M. Batet, A study on semantic similarity and its application to clustering: Enabling the classification of textual data, VDM Verlag Dr. Müller, 2012.

[6] M. Batet, A. Erola, D. Sánchez, J. Castellà-Roca, Utility preserving query log anonymization via semantic microaggregation, Information Sciences, 242 (2013) 49-63.

[7] M. Batet, D. Sánchez, A. Valls, K. Gibert, Semantic similarity estimation from multiple ontologies, Applied Intelligence, 38 (2013) 29-44.

[8] M. Batet, A. Valls, K. Gibert, Performance of Ontology-Based Semantic Similarities in Clustering, in: Artificial Intelligence and Soft Computing, 10th International Conference, ICAISC 2010, Part I, Springer-Verlag, Zakopane. Poland, 2010, pp. 281–288.

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

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

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

[12] G. Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction, in: Biennial GSCL Conference, Tübingen: Gunter Narr Verlag, Germany, 2009, pp. 31-40.

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

[14] K.W. Church, P. Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, 16 (1990) 22-29.

[15] C.E. Crangle, M. Perreau-Guimaraes, P. Suppes, Structural Similarities between Brain and Linguistic Data Provide Evidence of Semantic Relations in the Brain, PLoS ONE, 8 (2013) e65366.

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

[17] V.N. Garla, C. Brandt, Semantic similarity in the biomedical domain: an evaluation across knowledge sources, BMC Bioinformatics, 13 (2012) 261.

[18] T. Grego, F.M. Couto, Enhancement of Chemical Entity Identification in Text Using Semantic Similarity Validation, PLoS ONE, 8 (2013) e62984.

[19] P.H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Briefings in Bioinformatics, 13 (2012) 569-585.

[20] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies, Bioinformatics, (in press) (2013).

[21] S. Hoban, C. Vernesi, Challenges in global biodiversity conservation and solutions that cross sociology, politics, economics and ecology, Biology letters, 8 (2012) 897-899.

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

[23] M.C. Lange, D.G. Lemay, J.B. German, A multi-ontology framework to guide agriculture and food towards diet and health, Journal of the science of food and agriculture, 87 (2007) 1427–1434.

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

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

[26] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of Biomedical Informatics, 46 (2013) 294–303.

[27] S. Martínez, A. Valls, D. Sánchez, Semantically-grounded construction of centroids for datasets with textual attributes, Knowledge-Based Systems, 35 (2012) 160-172.

[28] B.T. McInnes, T. Pedersen, Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text, Journal of Biomedical Informatics, (2013) (in press).

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

[30] N.F. Noy, N.H. Shah, P.L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D.L. Rubin, M.-A. Storey, C.G. Chute, M.A. Musen, BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Research, 37 (2009) W170-W173.

[31] S. Pakhomov, B. Mcinnes, T. Adams, Y. Liu, T. Pedersen, G.B. Melton, Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study, in: Annual Symposium of the American Medical Association, Washington, D.C., 2010, pp. 572–576.

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

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

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

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

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

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

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

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

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

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

[42] D. Sánchez, M. Batet, A Semantic Similarity Method Based on Information Content Exploiting Multiple Ontologies, Expert Systems with Applications, 40 (2013) 1393–1399.

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

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

[45] D. Sánchez, M. Batet, A. Viejo, Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security, 8 (2013) 853-862.

[46] D. Sánchez, M. Batet, A. Viejo, Minimizing the disclosure risk of semantic correlations in document sanitization, Information Sciences, 249 (2013) 110-123.

[47] D. Sánchez, J. Castellà-Roca, A. Viejo, Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines, Information Sciences, 218 (2013) 17-30.

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

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

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

[51] K. Saruladha, G. Aghila, A. Bhuvaneswary, Information content based semantic similarity for cross ontological concepts, International Jornal of Engineering Science and Technology, 3 (2011) 5132 - 5140.

[52] N. Seco, T. Veale, J. Hayes, 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, IOS Press, Valencia, Spain, 2004, pp. 1089-1090.

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

[54] M.-F. Sy, S. Ranwez, J. Montmain, A. Regnault, M. Crampes, V. Ranwez, User centered and ontology based information retrieval system for life sciences, BMC Bioinformatics, 13 (2012) S4.

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

[56] C. Vicient, D. Sánchez, A. Moreno, An automatic approach for ontology-based feature extraction from heterogeneous textual resources, Engineering Applications of Artificial Intelligence, 26 (2013) 1092-1106.

[57] X. Wu, E. Pang, K. Lin, Z.-M. Pei, Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method, PLoS ONE, 8 (2013) e66745.

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

[59] Z. Zhou, Y. Wang, J. Gu, 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, IEEE Computer Society, Sanya, Hainan Island, China, 2008, pp. 85-89.