Journal of Biomedical Informatics 44 (2011) 118–125

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An ontology-based measure to compute semantic similarity in biomedicine

Montserrat Batet *, David Sánchez, Aida Valls

Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) Research Group, Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain

ARTICLE INFO

Article history:
Available online 15 September 2010

Keywords:
Semantic similarity
Ontologies
SNOMED CT
Biomedicine
Data mining

ABSTRACT

Proper understanding of textual data requires the exploitation and integration of unstructured and heterogeneous clinical sources, healthcare records or scientific literature, which are fundamental aspects in clinical and translational research. The determination of semantic similarity between word pairs is an important component of text understanding that enables the processing, classification and structuring of textual resources. In the past, several approaches for assessing word similarity by exploiting different knowledge sources (ontologies, thesauri, domain corpora, etc.) have been proposed. Some of these measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data or from medical ontologies (such as MeSH or SNOMED CT). In this paper, these approaches are introduced and analyzed in order to determine their advantages and limitations with respect to the considered knowledge bases. After that, a new measure based on the exploitation of the taxonomical structure of a biomedical ontology is proposed. Using SNOMED CT as the input ontology, the accuracy of our proposal is evaluated and compared against other approaches according to a standard benchmark of manually ranked medical terms. The correlation between the results of the evaluated measures and the human experts' ratings shows that our proposal outperforms most of the previous measures avoiding, at the same time, some of their limitations.

1. Introduction

In the last few years, the amount of clinical data that is electronically available has increased rapidly. Digitized patient health records and the vast amount of medical and scientific documents in digital libraries have become valuable resources for clinical and translational research. However, as translational research processes involve real world entities (such as patients) and events (such as patients' visits) whose associated data are mainly stored as documents (e.g., visit outcomes, empirical observations, worksheets, etc.) [1], most of the resulting information sources are presented in unprocessed and heterogeneous textual formats. Semantic technologies play an important role in this context enabling a proper interpretation of this information.

The determination of the *semantic similarity* between words constitutes a pillar of text understanding, being successfully applied in many natural language processing tasks such as word-sense disambiguation [2,3], document categorization or clustering [4,5], word spelling correction [6], automatic language translation [4], ontology learning [7] or information retrieval [8–10]. In the biomedical field, similarity computation can improve the performance of information retrieval from biomedical sources [11,10] and may ease the integration of heterogeneous clinical data [12].

Semantic similarity computes the likeness between words, understood as the degree of taxonomical proximity. For example, *bronchitis* and *flu* are similar because both are disorders of the respiratory system. However, words can also be related in non-taxonomical ways (e.g., *diuretics* help in the treatment of *hypertension*); in this more general case, one talks about *semantic relatedness*. In both sets of cases, they are based on the evaluation of the semantic evidence observed in a knowledge source (such as ontologies or domain corpora). According to the type of domain knowledge exploited, different families of functions can be identified: those based on the taxonomical structure of an ontology (discussed in Section 2), those relying on the information content (IC) of concepts (reviewed in Section 3) and those exploiting the amount of co-occurrences between word contexts (detailed in Section 4).

From a domain-independent point of view, these approaches provide accurate results when relying on large and general-purpose knowledge sources such as WordNet [13] and tagged corpora such as SemCor [14]. However, these measures perform poorly with biomedical terms due to the limited coverage of specialized domains [15] in the knowledge models. Fortunately, there are a

* Corresponding author. Address: Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) Research Group, Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain. Fax: +34 977559710.

E-mail address: montserrat.batet@urv.cat (M. Batet).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.09.002

number of relevant biomedical ontologies, knowledge repositories and structured vocabularies that model and organize concepts in a comprehensive manner. Well-known examples are MeSH (Medical Subject Headings) for indexing literature, the ICD taxonomy (International Classification of Diseases) for recording causes of death and diseases, and SNOMED CT. Several authors [11,16,17] have applied some of the classical similarity computation paradigms to medical data by exploiting SNOMED CT and/or clinical data. While some authors compared different approaches for similarity computation using SNOMED CT as knowledge source, evaluating them over particular datasets [18–20], or in the context of a concrete application, such as document clustering [5,21], some other authors exploited the MeSH ontology to compute the similarity assessment between words [10,18,22,23,24].

In this paper, we first review and analyze the measures for similarity/relatedness computation commonly referenced in the literature, with details of their adaptation to the biomedical domain. We review each family of measures to identify their advantages and limitations under the dimensions of expected accuracy, computational complexity, dependency on knowledge sources (size and pre-processing) and parameter tuning. In order to overcome some of the problems identified in this study, we present a new measure based on the exploitation of all the taxonomical knowledge regarding the compared concepts. Finally, the paper evaluates and compares the results obtained by our measure against those reported by other similarity functions when applied to the biomedical domain. The results show that our proposal provides a high accuracy without having some of the limitations identified on other measures.

The rest of the paper is organized as follows. Sections 2–4 present and analyze similarity measures belonging to the taxonomy-based, IC-based and context vector-based paradigms. Section 5 presents our similarity measure and its main advantages. Section 6 evaluates it using SNOMED CT as the domain ontology, and compares it against the analyzed measures. Section 7 analyzes and discusses the results. The final section presents the conclusions.

2. Similarity measures based on the taxonomical structure

The first family of measures exploits the geometrical model provided by concept hierarchies. Domain knowledge is explicitly modeled in a machine-readable language which formalizes domain concepts using a common terminology and represents taxonomic and non-taxonomical relationships via semantic links. In this case, the basis to compute concept resemblance is the inter-link distance.

In a taxonomy, the simplest way to estimate the distance (*dis*) between two concepts c_1 and c_2 is by calculating the shortest *path length* (*PL*, i.e., the minimum number of links) connecting these concepts [22]

$$dis_{PL}(c_1,c_2) = \min \text{ number of taxonomical edges connecting } c_1 \text{ and } c_2 \quad (1)$$

Several variations of this measure such as the one presented by Wu and Palmer [25] (*W&P*) have been developed. Considering that the similarity (*sim*) between a pair of concepts in an upper level of the taxonomy is smaller than the similarity between a pair in a lower level, they proposed a path-based measure that also takes into account the depth of the concepts in the hierarchy

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

where N_1 and N_2 are the number of *is-a* links from c_1 and c_2, respectively, to their least common subsumer (LCS), and N_3 is the number of *is-a* links from the LCS to the root of the ontology. It ranges from 1 (for identical concepts) to 0.

Leacock and Chodorow [26] (*L&C*) also proposed a measure that considers both the shortest path between two concepts (in fact, the number of nodes N_p from c_1 to c_2 including themselves) and the maximum depth D of the taxonomy

$$sim_{L\&C}(c_1,c_2) = -\log(N_p/2D) \quad (3)$$

Li et al. [27] proposed a similarity measure that combines the shortest path length and the depth of the ontology evaluated in a non-linear fashion

$$sim_{Li}(c_1,c_2)_{Li} = e^{-\alpha path(c_1,c_2)}\cdot\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}} \quad (4)$$

where $path(c_1, c_2)$ is the shortest path length between two concepts, h is the minimum depth of the LCS in the hierarchy and $\alpha \geqslant 0$ and $\beta > 0$ are parameters scaling the contribution of the shortest path length and depth, respectively. Based on benchmark data, the optimal parameters for the measure were $\alpha = 0.2$; $\beta = 0.6$.

Choi and Kim [28] also proposed a similarity measure applied to the Yahoo! category tree for solving the problem of topic distillation. The measure is computed according to the difference on the depth levels of two concepts and the distance of the shortest path between them

$$sim_{CK}(c_1,c_2) = \frac{MAX\_PATH - path(c_1,c_2)}{MAX\_PATH}\times\frac{MAX\_LEVEL - diff\_level(c_1,c_2)}{MAX\_LEVEL} \quad (5)$$

Al-Mubaid and Nguyen [16] proposed a cluster-based measure that combines *path length* and *common specificity*. They define clusters for each of the branches in the hierarchy with respect to the root node. The common specificity is used to state that lower level pairs of concept nodes should be considered more similar than higher level pairs. The common specificity of two concepts is measured by subtracting the depth of their LCS from the depth D_c of the cluster

$$CSpec(c_1,c_2) = D_c - depth(LCS(c_1,c_2)) \quad (6)$$

So, the smaller the common specificity of two concept nodes, the more the information they share, and thus, the more similar they are. The proposed distance measure (*sem*) is defined as follows:

$$dis_{sem}(c_1,c_2) = \log((path(c_1,c_2)-1)^{\alpha}\times(CSpec)^{\beta}+k) \quad (7)$$

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

The advantage of measures based on the *taxonomy structure exploitation* paradigm is that they only use an ontology as background knowledge (i.e., no corpus with domain data is needed). However, their main problem is that they heavily depend on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology [29]. Moreover, as most of them base the similarity assessment only on the *minimum path*, they may omit a large amount of taxonomical knowledge available in the ontology for the given pair of concepts (e.g., the full set of common and non-common ancestors). As a consequence, these measures are typically surpassed by approaches based on exploiting additional semantic evidence inferred from the information distribution of a domain corpus, such as IC- or context-vector-based ones [30]. Finally, it is worth noting that the presence of an *is-a* link between two concepts gives evidence of a taxonomic relationship but not of the degree of their semantic similarity,

because all individual links have the same length and, in consequence, represent uniform distances [31].

From a domain-independent point of view, the introduced approaches rely on large and general purpose repositories such as WordNet [13] (a freely available lexical database that describes, structures and links via taxonomic and non-taxonomic semantic pointers more than 100,000 general English concepts). WordNet's large taxonomy, with a relatively homogeneous distribution of semantic links and good inter-domain coverage is the ideal environment for applying these measures [32]. However, due to the limited WordNet coverage of biomedical terms [15], the accuracy of similarity assessments for medical terms is poor [11].

For this reason, Pedersen et al. [11] and Al-Mubaid and Nguyen [16,17] have adapted these measures to the biomedical domain by exploiting SNOMED CT. SNOMED CT (*Systematized Nomenclature of Medicine, Clinical Terms*) is an ontological/terminological resource distributed as part of the UMLS. It is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerized physician order entry, disease surveillance, image indexing and consumer health information services. It contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into 18 overlapping hierarchies: clinical findings, procedures, observable entities, body structures, organisms, substances, pharmaceutical products, specimens, physical forces, physical objects, events, geographical environments, social contexts, linkage concepts, qualifier values, special concepts, record artifacts and staging and scales. Each concept may belong to one or more of these hierarchies by multiple inheritance (e.g., euthanasia is an event and a procedure), or it may inherit from multiple concepts within one of these hierarchies. Concepts are linked with approximately 1.36 million relationships. In such a complete domain ontology, *is-a* relationships have been exploited to estimate term similarity, even though much of the taxonomical knowledge explicitly modeled is still unexploited.

3. IC-based similarity measures

Information content (IC) measures the amount of information provided by a given term based on its probability of appearance in a corpus. Formally, the IC of a concept c is the inverse of its probability of occurrence, $p(c)$ (8). In this manner, infrequent words are considered as more informative than common ones

$$IC(c) = -\log p(c) \tag{8}$$

Based on this premise, Resnik [33] presented a seminal work in which the similarity between a pair of concepts (c_1 and c_2) is estimated as the amount of taxonomical information they share. In a taxonomy, this information is represented by the *least common subsumer* of both terms ($LCS(c_1, c_2)$), which is the most specific taxonomical ancestor common to c_1 and c_2 in a given ontology (9). This gives an indication of the amount of information that the two concepts share. The more specific the subsumer is (higher IC), the more similar the terms are

$$sim_{res}(c_1, c_2) = IC(LCS(c_1, c_2)) \tag{9}$$

The most commonly used extensions of Resnik's measure are those of Lin [34] and Jiang and Conrath [32].

Lin similarity depends on the relation between the information content of the LCS of two concepts and the sum of the information content of the individual concepts

$$sim_{lin}(c_1, c_2) = \frac{2 \times sim_{res}(c_1, c_2)}{(IC(c_1) + IC(c_2))} \tag{10}$$

Jiang and Conrath subtract the information content of the LCS from the sum of the information content of the individual concepts

$$dis_{jcn}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2 \times sim_{res}(c_1, c_2) \tag{11}$$

Note that this is a dissimilarity measure because the more different the terms are, the higher the difference between their IC and the IC of their LCS will be.

In order to obtain reliable results using these measures, the way in which the probability $p(c)$ is computed is crucial. Authors obtained near baseline results (compared to human judgments [35]) when obtaining the LCS from WordNet and estimating word frequencies from SemCor [14] (a WordNet-based semantically tagged text consisting in 100 passages from the Brown Corpus). Considering that the tagging scheme of SemCor was based on the list of concepts covered by WordNet 1.6 and that WordNet is also used by these measures to extract the LCS, frequency distribution for each concept is very precise. In fact, using this corpus as background, these measures outperform path length-based ones [32]. The drawback is the small size and high data sparseness of background data (i.e., the fact that the available data is not enough to extract valid conclusions about domain information distribution), due to the need of manually tagging the sense for each word in the corpus. As a result, less than 13% of the word senses available in the latest version of WordNet (3.0) appear in the corpus.

The coherence of the IC computation with respect to the taxonomical structure is the other aspect that should be ensured in order to maintain the consistency of the similarity computation. Resnik-based measures explicitly introduce the premise that the subsumer's IC must be lower than its specializations. To guarantee this property, Resnik [33] proposed to calculate the probability of a concept as the sum of the individual occurrences of all the concepts which are subsumed by it

$$p(c) = \sum_{n \in specializations(c)} \frac{count(n)}{N} \tag{12}$$

where $specializations(c)$ are the set of terms subsumed by concept c, and N is the total number of concepts observed in the corpus.

However, this approach forces the recursive computation of all the appearances of the subsumed terms to obtain the subsumer's IC. If either the taxonomy or the corpus changes, re-computations of the affected branches are needed, hampering the scalability of the solution. Moreover, the background taxonomy must be as complete as possible (i.e., it should include most of the specializations of a specific concept) in order to provide reliable results. Partial taxonomies with a limited scope may not be suitable for this purpose.

Finally, using a general purpose corpus such as SemCor to estimate the information distribution also hampers the performance of biomedical similarity assessments, due to its reduced coverage of biomedical terms. Considering that medicine is a large and complex domain which is rich in synonymy and semantically similar or related concepts, these general purpose repositories are not ideal [21]. Pedersen et al. [11] adapted IC-based measures to the biomedical domain by exploiting SNOMED CT taxonomy and the Mayo Clinic Corpus of Clinical Notes as a domain corpus. As stated by the authors, the Mayo Clinic Corpus consists of 1,000,000 clinical notes collected over the year 2003 which cover a variety of major medical specialties at the Mayo Clinic. Clinical notes have a number of specific characteristics that are not found in other types of discourse, such as news articles or scientific medical articles found in MEDLINE. They are generated in the process of treating a patient and contain the record of the patient–physician encounter. Notes were transcribed by trained personnel and structured according to the reasons, history, diagnostic, medications and other administrative information. Patient's history, reason for visit and diagnostic related notes were used as the domain-specific data corpus from which IC-based measures can be computed. Unfortunately, big and detailed corpora are not typically available for

many domains due to the cost of compiling, structuring and processing such an amount of information and, in some cases, due to security issues regarding the private nature of personal data. As a result, IC-based measures may be compromised by the availability of enough suitable data.

4. Context vector relatedness measures

The third family of measures computes semantic likeness by exploiting the hypothesis that words are similar if their contexts are similar [36]. In particular, these measures construct co-occurrence vectors that represent the contextual profile of concepts (*context vectors*). Context vectors are built by extracting contextual words (within a fixed window of context) from a corpus of textual documents covering the evaluated concepts [37]. These vectors capture a more general sense of concept likeness, not necessarily reduced to taxonomical similarity but also to inter-concept relatedness.

The semantic relatedness of two concepts c_1 and c_2 is computed as the cosine of the angle between their context vectors [30]

$$rel_{vector}(c_1, c_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|v_1| \cdot |v_2|} \quad (13)$$

where $\vec{v}_1$ and $\vec{v}_2$ are the context vectors corresponding to c_1 and c_2, respectively.

In a domain independent setting, Patwardhan and Pedersen [30] created vectors from term glosses extracted from WordNet, calling them *gloss vectors*. Glosses are brief and explanatory notes about the meaning of a particular word sense. They are manually created by knowledge experts, so, they represent the ideal context of a concept. In fact, glosses would likely contain terms that may help to distinguish concepts better than text extracted from a generic corpus of raw text. Considering WordNet glosses as a corpus of contexts one obtains about 1.4 million words, which should be processed in order to create the context vectors (introducing a noticeable computational cost). As a result, the gloss vector measure was able to obtain the highest correlation with regards to human judgments in several domain independent benchmarks [30]. Nevertheless, as stated by Pedersen et al. [11], the quality of the assessment heavily depends on the tuning, nature and size of the corpus from which the context vectors are created.

In the biomedical field, the same authors adapted their measure by using the mentioned Mayo Clinic Corpus of Clinical Notes and the Mayo Clinic Thesaurus to extract context words and term descriptions, respectively, within a context window of one line of text. The Mayo Clinic thesaurus is a source of clinical problem descriptions that have been collected in the Mayo Clinic [11] (i.e., the equivalent to WordNet glosses). It contains 16 million diagnostic phrases expressed in natural language classified in over 21,000 categories. Authors took these phrases to generate quasi-definitions (term descriptions) for terms found in SNOMED CT, after a pre-processing stage aimed to reduce noise and redundancy of natural-language text. The context words of the terms found in the descriptions extracted from the Clinical Notes repository were aggregated to get the context vector of a concept. Similarly to IC-based measures, corpora availability and suitability are the main problems that hamper the applicability of these measures.

5. A new measure to compute the semantic similarity

From the study of similarity measures described in previous sections and considering how they have been applied to the biomedical domain, we can extract the following conclusions. From the applicability point of view, path-based measures are the most adequate ones. As they only exploit the geometrical model of the ontology, no pre-calculus or pre-processing is needed, which makes them more computationally efficient. However, due to their simplicity, they do not capture enough semantic evidence to provide assessments as reliable as other types of measures (as it will be shown in Section 6).

On the contrary, measures based on IC and context vectors require additional domain data in order to provide more accurate assessments in comparison to path-based measures. The fact that corpora consist of unstructured or slightly structured natural-language text implies that a certain degree of pre-processing is needed to extract implicit semantic evidence and to provide accurate results. In general, the more the pre-processing of the corpus is performed (in order to reduce noise or language ambiguity), the more accurate the results can potentially be. In fact, the size of corpora needed to provide good assessments is so big (millions of words) that their pre-processing introduces a serious computational burden. Moreover, in the biomedical field, the availability of a big and heterogeneous corpus of clinical data is especially problematic due to the sensitivity of patient data, which may result in data sparseness problems [38]. Summarizing, even though a corpus-based approach may lead to more accurate results, their dependency on data availability, suitability and pre-processing hampers their real applicability.

Taking these conclusions into account, we propose a new similarity measure that can achieve a level of accuracy similar to corpus-based approaches but retaining the low computational complexity and lack of constraints of path-based measures (i.e., no domain corpus is needed).

Analyzing the basic hypothesis of path-based methods, we can notice that these measures consider the minimum path length between a concept c_1 and a concept c_2, which is the sum of taxonomical links between each of the concepts and their LCS. The path is composed, in addition to the LCS, of nodes corresponding to non-shared superconcepts (i.e., subsumers of the evaluated terms), which are taken as an indication of distance. However, if one or both concepts inherit from several *is-a* hierarchies, all possible paths between the two concepts are calculated, but only the shortest one is kept. In consequence, the resulting path length does not completely measure the total amount of non-common superconcepts modeled in the ontology (i.e., subsumers of a concept). Due to this reason, for complex and large taxonomies, such as SNOMED CT, covering thousands of interrelated concepts included in several overlapping hierarchies, and an extensive use of multiple inheritance (i.e., a concept is subsumed by several superconcepts), path-based measures waste a great amount of explicit knowledge. So, it seems reasonable that a measure that takes into account all the available taxonomical evidence (i.e., all the superconcepts) regarding the evaluated concepts (and not only the minimum path) could provide more accurate assessments.

In our proposal, in order to capture as much semantic evidence as possible in the case of multiple inheritances, we take all the superconcepts belonging to all the possible taxonomical paths connecting the evaluated terms. That is, for a given pair of distinct concepts, we consider the concepts and their *complete* set of non-shared superconcepts as an indication of their distance. By considering concepts themselves in conjunction with the set of non-common superconcepts we are able to calculate the similarity for a pair of concepts that are siblings of an immediate superclass (i.e., they share their complete sets of superconcepts).

However, by considering only non-shared knowledge, we are not able to distinguish concepts with very few or even no superconcepts in common from others with more communal information. For example, as shown in Fig. 1, the number of non-common superconcepts for the pair (c_1, c_2) and for the concepts (c_3, c_4) is equal; but, it makes sense that the distance between c_1 and c_2 is lower than the distance between c_3 and c_4 due to the

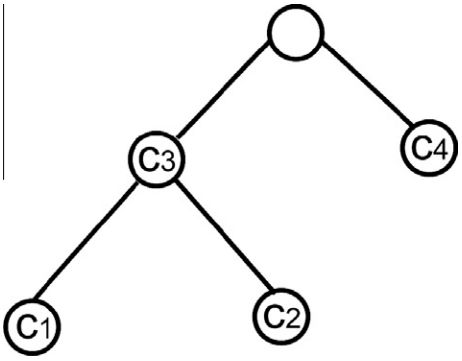

Fig. 1. Taxonomy example.

higher amount of shared superconcepts of the pair (c_1, c_2). This is also related to the assumption formulated by some authors [25] who consider that pairs of concepts belonging to an upper level of the taxonomy (i.e., they share few superconcepts) should be less similar than those in a lower level (i.e., they have more superconcepts in common).

In order to take into account the amount of common information between a pair of concepts, we define our measure as the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge. As a result, this definition introduces a desired penalization to those cases in which the number of shared superconcepts is small. In the example of Fig. 1, as the number of common superconcepts between (c_1, c_2) is 2 and between (c_3, c_4) is only 1 and, in both cases, the number of non-common superconcepts plus the non-equal concept pair is 2 (i.e., only the concepts themselves), the distance between (c_1, c_2) will be smaller than between (c_3, c_4), being $2/(2+2) = 0.5$ and $2/(1+2) = 0.66$, respectively.

Finally, considering that shared and non-shared knowledge explicitly retrieved from a repository for a concept pair is not linear to their similarity/distance [39], we introduce the inverted logarithm function to smooth the assessments and to transform the function into a similarity. In fact, in Ref. [16] it is argued that a non-linear approach is the optimum one for combining semantic features. The final similarity measure is presented in Eq. (14).

Let us define the full concept hierarchy or taxonomy (H^C) of concepts (C) of an ontology as a transitive *is-a* relation $H^C \in C \times C$, and we define $T(c_i) = \{c_j \in C \mid c_j \text{ is superconcept of } c_i\} \cup \{c_i\}$ as the union of the ancestors of the concept c_i and c_i itself.

Then, the similarity measure between two concepts is defined as:

$$sim(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \quad (14)$$

Note that, for the case of a concept being compared with itself, the set of differential elements (numerator) will be zero, resulting in an infinitely large similarity value. In order to avoid such an infinite value, it must be checked that the two concepts being compared are distinct before applying our similarity measure.

Computationally, this measure retains the simplicity of path-based approaches, being much simpler than the calculus needed to estimate the information distribution in a corpus or to pre-process it. Next, we will evaluate its accuracy in the biomedical domain with respect to other measures.

6. Evaluation

The usual way of evaluating the accuracy of similarity measures is based on using a set of pairs of words whose similarity has been assessed by a group of human experts. Computing the correlation between the similarity ratings obtained by a computerized approach with respect to the human judgments, one is able to obtain a quantitative value about the quality of the similarity estimation. This enables an objective comparison among different measures. In a general setting, the most commonly used benchmarks are the Miller and Charles [35] and Rubenstein and Goodenough [40] sets, which are composed of manually rated lists of domain-independent pairs of words.

For the biomedical domain, several authors [19,20,10] have created *ad hoc* datasets to evaluate their approaches, framed on concrete research projects or oriented to particular ontologies. This hampers their applicability as general purpose domain benchmarks. Pedersen et al. [11] stated the necessity of having objectively scored datasets that could be used as a direct means of evaluation in the biomedical domain. Thus, they created, in collaboration with Mayo Clinic experts, a benchmark referring to medical disorders. The similarity between term pairs was assessed by a set of nine medical coders who were aware of the notion of semantic similarity and a group of three physicians who were experts in the area of rheumatology. After a normalization process, a final set of 30 word pairs were rated with the average of the similarity values provided by the experts in a scale between 1 and 4 (see Table 1). The correlation between the physicians was 0.68, whereas the correlation between medical coders achieved a value of 0.78.

Pedersen et al. [11] used that benchmark to evaluate most of the measures based on path length and information content, and their own context vector measure, by exploiting SNOMED CT as the domain ontology[1] and the Mayo Clinical Corpus and Thesaurus as corpora. Al-Mubaid and Nguyen [16] also used that benchmark and SNOMED CT to evaluate other path-based measures also considered in this paper, including their own proposal (*sem*); in this case, results were only compared against coders' ratings because they considered them to be more reliable than physicians' judgments. Note that whereas human ratings measure similarity, some of the evaluated measures compute distance. In these cases a linear transformation was performed.

In order to enable an objective comparison between our proposal and other measures in the biomedical domain, we have also used the benchmark of Pedersen et al. and the SNOMED CT ontology to evaluate the accuracy of our measure. Correlation values obtained for our measure together with their *standard errors* and correlations reported by related works with respect to both sets of human experts are presented in Table 2. Note that, for the context vector measure, four different tests are reported, changing two of its most influential parameters: corpus size (1 million or 100,000 clinical notes) and corpus selection (considering only the diagnostic section of clinical notes or all the sections of the document).

7. Discussion

Analyzing the results presented in the previous section (Table 2) and considering the correlation values between human experts (0.68 for physicians and 0.78 for coders), it can be seen that path length-based measures offered a limited performance, with correlations smaller than 0.36 and 0.66, respectively. This shows that limited accuracy is obtained when estimating semantic similarity only from the minimum inter-link path. In complex domain ontologies, such as SNOMED CT, where multiple paths between concepts constructed from several overlapping taxonomies are available, this approach wastes a lot of explicitly available knowledge. In fact, the measure with the best accuracy (0.66 for coders)

[1] Note that the pair "*chronic obstructive pulmonary disease*" – "*lung infiltrates*" was excluded from the test as the latter term is not found in the SNOMED CT terminology.

Table 1
Set of 30 medical term pairs with averaged experts' similarity scores (extracted from [11]).

Term 1	Term 2	Physician ratings (averaged)	Coder ratings (averaged)
Renal failure	Kidney failure	4.0	4.0
Heart	Myocardium	3.3	3.0
Stroke	Infarct	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Delusion	Schizophrenia	3.0	2.2
Congestive heart failure	Pulmonary edema	3.0	1.4
Metastasis	Adenocarcinoma	2.7	1.8
Calcification	Stenosis	2.7	2.0
Diarrhea	Stomach cramps	2.3	1.3
Mitral stenosis	Atrial fibrillation	2.3	1.3
Chronic obstructive pulmonary disease	*Lung infiltrates*	2.3	1.9
Rheumatoid arthritis	Lupus	2.0	1.1
Brain tumor	Intracranial hemorrhage	2.0	1.3
Carpal tunnel syndrome	Osteoarthritis	2.0	1.1
Diabetes mellitus	Hypertension	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Cortisone	Total knee replacement	1.7	1.0
Pulmonary embolus	Myocardial infarction	1.7	1.2
Pulmonary fibrosis	Lung cancer	1.7	1.4
Cholangiocarcinoma	Colonoscopy	1.3	1.0
Lymphoid hyperplasia	Laryngeal cancer	1.3	1.0
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Rectal polyp	Aorta	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Depression	Cellulitis	1.0	1.0
Varicose vein	Entire knee meniscus	1.0	1.0
Hyperlipidemia	Metastasis	1.0	1.0

Table 2
Correlation values obtained for each measure against ratings of physicians, coders and both. For our measure, standard errors are presented in parentheses.

Measure	Evaluated in Refs.	Physicians	Coders	Both
Path	[11]	0.36	0.51	0.48
Wu and Palmer	[16]	N/A	0.29	N/A
Leacock and Chodorow	[11]	0.35	0.50	0.47
Li et al.	[16]	N/A	0.37	N/A
Choi and Kim	[16]	N/A	0.15	N/A
sem	[16]	N/A	0.66	N/A
Resnik	[11]	0.45	0.62	0.55
Lin	[11]	0.60	0.75	0.69
Jiang and Conrath	[11]	0.45	0.62	0.55
Context vector (1 million notes, diagnostic section)	[11]	0.84	0.75	0.76
Context vector (1 million notes, all sections)	[11]	0.62	0.68	0.69
Context vector (100,000 notes, diagnostic section)	[11]	0.56	0.59	0.60
Context vector (100,000 notes, all sections)	[11]	0.41	0.53	0.51
Our measure	–	0.60 (±0.119)	0.79 (±0.07)	0.73 (±0.087)

is *sem*, which, as introduced in Section 2, based the assessment both in the path length and in the relative depth of the concepts. This captures more knowledge than measures based only on the absolute path and the global depth of the ontology.

With regards to IC-based measures, in general, they are able to improve the results of path-length approaches. The maximum correlations are 0.6 for the physicians and 0.75 for the coders. Moreover, the minimum correlation for coders is 0.62, which outperforms path length results (with the exception of *sem*). The fact of relying on high-quality domain corpora (i.e., clinical notes) allows complementing the taxonomical knowledge extracted from the ontology with additional semantic evidence, given by the distribution of the information of the concept in domain corpora. However, as stated in Section 3, the applicability of these measures is hampered by the dependency on the availability and adequacy of domain data with respect to the evaluated concepts.

For the context vector measure, four cases were presented by the authors, changing the corpus size and corpus selection. The correlation strongly depends on the amount and quality of the background corpus (with values between 0.51 and 0.76 considering the average of both sets of experts). The best accuracy (correlations of 0.84 for the physicians and 0.75 for the coders) is achieved under particular circumstances: 1 million notes involving only the diagnostic section. In this case, due to the fact that term definitions are extracted from high-quality corpora and due to the enormous size of the information sources, the obtained context vectors can adequately define the evaluated terms, enabling accurate estimates. However, for other corpus configurations, the accuracy of the measure decreases noticeably, at levels even below path-based approaches like *sem*. In fact, it drops to correlations of 0.41 for physicians and 0.53 for coders when 100,000 notes involving all sections are used.

Regarding the proposed semantic similarity, in order to measure the statistical significance of our results, we conducted the following analyses. On one hand, when dealing with a relatively low sample size, we cannot always be sure that correlation values are accurate or occurred by chance (i.e., being zero correlated in the worst case). To tackle this problem, we measured the *significance* of our correlation by computing the *p-value* for the correlation values, which tells the probability that the observed correlation occurred by chance. In the three cases, the *p-values* for our results were less than 0.001 (0.1% chance), which indicates that correlation values are statistically significant.

On the other hand, in order to measure the significance of the differences observed between the different correlation values, we also computed the *standard error* (SE) of our correlation values. We obtained a SE of 0.119 for physicians ratings (defining a correlation error range of 0.481–0.719), 0.07 for coders ratings (defining a correlation error range of 0.72–0.86) and 0.087 when the average of both ratings were considered (defining a correlation error range of 0.643–0.817) as indicated in Table 2.

Compared to other approaches, and considering the calculated SEs, the correlation values obtained by our measure for the evaluated benchmark are, in all cases, higher than those reported for path-based measures. It is particularly interesting to see how, being a pure ontology-based measure, our proposal reports higher correlations than some IC-based measures (concretely the ones defined by Resnik and Jiang and Conrath); only Lin's measure reports correlation values which fall within the error ranges defined by our SEs (even though our correlation values are higher). This shows that the exploitation of all the taxonomical knowledge available in the ontology provides comparable or even more semantic evidence than other approaches exploiting additional data sources. In fact, the set of common and non-common superconcepts considered by our proposal incorporates, in an indirect manner, evidence of all the possible taxonomical paths between concepts, relative depths of branches and the relative densities of the involved taxonomical branches. This knowledge is implicit in the proposed way to calculate the semantic similarity between concepts. As stated during the review of the related work, in other ontology-based approaches these semantic features are only partially considered, obtaining less accurate assessments. The context vector measure, however, offered a comparable (considering the error ranges) or even better correlation with regards to our approach, when the complete amount of data and/or the diagnostic notes were used. In fact, it reported the highest correlation value (0.84) for physicians when using all the diagnostic notes.

Our measure obtains correlations which are comparable to the correlation between human experts: 0.60 vs 0.68 in the case of physicians, and 0.79 vs 0.78 with respect to medical coders. Analyzing in detail the different correlation values obtained with respect to the physicians' and the coders' ratings, one can notice important differences between the similarity measures. In general, all the measures, except the context-vector-based ones, correlate better with coders than with physicians. On one hand, this is motivated by the amount of discrepancies observed in the physicians' ratings, which correlate lower than those of coders' (0.68 vs 0.78). On the other hand, the way in which human experts interpret concept likeness also influences the results. During the construction of the original data set, medical coders were requested to reproduce the classical Rubenstein and Goodenough [40], Miller and Charles [35] benchmarks in order to ensure that coders understood the instructions and the notion of similarity. However, physicians rated the pairs of concepts without pre-training or external influences. As a consequence, medical coders' ratings, which are trained in the use of hierarchical classifications, seem to reproduce better the concept of (taxonomic) *similarity* whereas physicians' ratings seem to represent a more general concept of (taxonomic and non-taxonomic) *relatedness*. These intuitions are coherent with the fact that the context vector measure estimates *relatedness*, whereas the other ontology-based measures estimate *similarity*.

In addition, for the tests with context vector measure, the data corpus used to create vectors was constructed by physicians of the same clinic; so, it is biased towards the way in which physicians interpret and formalize knowledge. As stated by Pedersen et al. [11], these clinical notes may reflect implicit relations between concepts which were taken into consideration during the ratings and which are not explicitly indicated in a more general ontology such as SNOMED CT. Again, it makes sense that all the similarity measures correlate better with the less biased coders' ratings. In contrast, the unique relatedness measure considered in this review (context vector), which exploits the data composed by the same type of professionals which rated the benchmark, behaves in the inverse manner.

Finally, we would like to further analyze the situation in which the context vector measures significantly surpass the correlation obtained with our new method. Correlation values higher than 0.6 (with respect to physicians) are obtained when a huge amount of data (1 million clinical notes) is used to create the vectors. One can see how the accuracy of the measure decreases when a narrower corpus is used. This dependency on the corpus size implies that the amount of processing needed to create the vectors from such an amount of data is not negligible. Moreover, the highest correlation is only obtained when using a particular subset of data, which corresponds to the descriptions of diagnostics and treatments. As stated by Pedersen et al. [11], this section contains more closely related terms than others which involve more noisy data. In consequence, as stated above, the choice, size and processing of the corpora used with the context vector measure is critical to achieve a good accuracy. This requires making a number of informed choices a priori in order for the measure to behave optimally for a concrete situation and domain.

On the contrary, our measure, which is based only on an ontology, is able to provide a comparatively high accuracy without any dependency on data availability and pre-processing (which would hamper its applicability) and, at the same time, retains the low computational complexity and lack of constraints of path-based measures. As any other ontology-based measure, the final accuracy will depend on the detail, completeness and coherency of taxonomical knowledge. Moreover, most of the improvements achieved by our approach are derived from the fact that similarity is estimated from the total set of subsumer concepts considering the different taxonomical hierarchies. If the input ontology offers little taxonomical detail or does not consider multiple inheritances between concepts, the accuracy improvements of our approach with respect to measures based on the minimum path are likely to be less noticeable. Fortunately, large, broad and widely used structured knowledge sources are ideal backgrounds for our measure because concepts belong to multiple and detailed taxonomies, like those included in the UMLS repository or general purpose ones like WordNet.

After this study, we plan to evaluate our approach in other specific domains such as tourism, in which textual comprehension and ontologies play an important role when developing new intelligent services [41]. Furthermore, the exploitation of other non-taxonomic relationships available in repositories such as SNOMED CT (e.g., attributes or other relations that represent other characteristics of the concept) or WordNet (e.g., meronyms, holonomy, antonymy or related terms) [42] should also be studied since they could provide additional evidence about concept *relatedness*, without compromising the generality of the approach.

In addition, we are interested in applying these measures in unsupervised clustering methods. In this field, concepts are treated as categorical or nominal values without any semantic

interpretation of the terms. The use of a semantic similarity measure to compare objects permits the inclusion of background knowledge provided by specific domain ontologies. A first attempt to include semantic measures has been done [43] and further studies are in progress.

8. Conclusions

Considering that large ontologies like SNOMED CT offer detailed taxonomic knowledge which is not exploited by path-based similarity measures, we proposed a measure that compiles as much taxonomic knowledge as available. As a result, it retains the simplicity and lack of corpora dependency of pure ontology-based approaches, but it improves their accuracy by exploiting additional semantic evidence. The evaluation sustained this idea, showing that our approach is able to outperform all previous path-based measures. It has also obtained comparable or better results than most of the measures based on IC and some tests based on context vectors. The latter, however, are hampered by their dependency on domain data availability, corpora pre-processing and parameter tuning, which require considerable human intervention. The fact that our measure provided the highest correlation value with respect to medical coders is particularly interesting, showing the reliability of the results in relation to the judgments of domain knowledge experts.

Acknowledgments

This work has been partially supported by the Universitat Rovira i Virgili (2009AIRE-04), the Spanish Ministry of Science and Innovation (DAMASK project, *Data mining algorithms with semantic knowledge*, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan). Montserrat Batet is also supported by a research grant provided by the Universitat Rovira i Virgili. The authors also acknowledge the review efforts of Dr. John Nealon and Dr. Susanne Greenwood from Oxford Brookes University and Joseph Soniran.

References

[1] Mirhaji P, Zhu M, Vagnoni M, Bernstam E, Zhang J, Smith J. Ontology driven integration platform for clinical and translational research. BMC Bioinf 2009;10(S-2). Available from: http://dx.doi.org/10.1186/1471-2105-10-S2-S2.
[2] Resnik P. Semantic similarity in a taxonomy. An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res 1999;11:95–130.
[3] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the fourth international conference on intelligent text processing and computational linguistics. Mexico City, Mexico; 2003. p. 241–57.
[4] Cilibrasi R, Vitányi PM. The google similarity distance. IEEE Trans Knowl Data Eng 2006;19(3):370–83.
[5] Aseervatham S, Bennani Y. Semi-structured document categorization with a semantic kernel. Pattern Recognit 2009;42(9):2067–76.
[6] Budanitsky A, Hirst G. Evaluating WordNet-based measures of semantic distance. Comput Linguist 2006;32(1):13–47.
[7] Sánchez D. Domain ontology learning from the web. VDM Verlag; 2008.
[8] Lee J, Kim M, Lee Y. Information retrieval based on conceptual distance in is-a hierarchies. J Doc 1993;49(2):188–207.
[9] Ratprasartporn N, Po J, Cakmak A, Bani-Ahmad S, Ozsoyoglu G. Context-based literature digital collection search. Int J Very Large Data Bases 2009;18(1):277–301.
[10] Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EGM, Milios EE. Information retrieval by semantic similarity. Int J Semantic Web Inf Syst 2006;2(3):55–73.
[11] Pedersen T, Pakhomov S, Patwardhan S, Chute C. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40:288–99.
[12] Sugumaran V, Storey VC. Ontologies for conceptual modeling: their creation, use, and management. Data Knowl Eng 2002;42:251–71.
[13] Fellbaum C. WordNet: an electronic lexical database. Cambridge, Massachusetts: MIT Press; 1998.
[14] Miller G, Leacock C, Tengi R, Bunker RT. A semantic concordance. In: Proceedings of ARPA workshop on human language technology – association for computational linguistics. Princeton, New Jersey, USA; 1993. p. 303–08.
[15] Burgun A, Bodenreider O. Comparing terms, concepts and semantic classes in WordNet and the unified medical language system. In: Proceedings of the NAACL 2001 (North American Association for Computational Linguistics) Workshop, WordNet and other lexical resources: applications, extensions and customizations. Pittsburgh, PA; 2001. p. 77–82.
[16] Al-Mubaid H, Nguyen HA. A cluster-based approach for semantic similarity in the biomedical domain. In: Conference proceedings of the IEEE engineering in medicine and biology society. New York, USA; 2006; p. 2713–7.
[17] Al-Mubaid H, Nguyen HA. Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Trans Syst Man Cybern 2009;39(4):389–98.
[18] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. J Biomed Inform 2004;37(2):77–85.
[19] Lee WN, Shah N, Sundlass K, Musen M. Comparison of ontology-based semantic-similarity measures. In: AMIA annual symposium proceedings. USA; 2008. p. 384–88.
[20] Matar Y, Egyed-Zsugmond E, Lajmi S. KWSim: concepts similarity measure. In: Proceedings of conférence en recherche d'Information et applications (CORIA08). France; 2008. p. 475–82.
[21] Melton G, Parsons S, Morrison F, Rothschild A, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006;39(6):697–705.
[22] Rada R, Mili H, Bichnell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;9(1):17–30.
[23] Petrakis GMP, Varelas G, Hliaoutakis A, Raftopoulou R. X-Similarity: computing semantic similarity between concepts from different ontologies. J Digit Inf Manage 2006;4:233–7.
[24] Pirró G. A semantic similarity metric combining features and intrinsic information content. Data Knowl Eng 2009;68:1289–308.
[25] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting of the association for computational linguistics. New Mexico, USA: Association for Computational Linguistics; 1994. p. 133–38.
[26] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In: WordNet: an electronic lexical database. MIT Press; 1998, p. 265–283.
[27] Li Y, Bandar ZA, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 2003;15(4):871–82.
[28] Choi I, Kim M. Topic distillation using hierarchy concept tree. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada; 2003. p. 371–72.
[29] Cimiano P. Ontology learning and population from text. Algorithms, evaluation and applications. Springer-Verlag; 2006.
[30] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006, workshop on making sense of sense: bringing computational linguistics and psycholinguistics together. Trento, Italy; 2006. p. 1–8.
[31] Bollegala D, Matsuo Y, Ishizuka M. WebSim: a web-based semantic similarity measure. In: The 21st annual conference of the Japanese society for artificial intelligence (JSAI2007). Miyazaki, Japan; 2007. p. 757–66.
[32] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics (ROCLING X). Taiwan; 1997. p. 19–33.
[33] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence (IJCAI 95). Montreal, Canada; 1995. p. 448–53.
[34] Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning (ICML98). Madison, Wisconsin, USA; 1998. p. 296–304.
[35] Miller G, Charles W. Contextual correlates of semantic similarity. Lang Cogn Process 1991;6(1):1–28.
[36] Harris Z. Distributional structure. In: Katz JJ, editor. The philosophy of linguistics. New York: Oxford University Press; 1985. p. 26–47.
[37] Schutze H. Automatic word sense discrimination. Comput Linguist 1998;24(1):97–123.
[38] Brill E. Processing natural language without natural language processing. In: Proceedings of the 4th international conference on computational linguistics and intelligent text processing. Mexico City, Mexico; 2003. p. 360–69.
[39] Lemaire B, Denhière G. Effects of high-order co-occurrences on word semantic similarities. Curr Psychol Lett 2006;18(1):23. Available from: http://cpl.revues.org/document471.html.
[40] Rubenstein H, Goodenough JB. Contextual correlates of synonymy. Commun ACM 1965;8:627–33.
[41] Prantner K, Ding Y, Luger M, Yan Z. Tourism ontology and semantic management system: state-of-the-arts analysis. In: Proceedings of IADIS (International Association for Development of the Information Society) international conference WWW/Internet 2007. Vila Real, Portugal; 2007. p. 111–15.
[42] Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum Christiane, editor. WordNet: an electronic lexical database. MIT Press; 1998. p. 305–32 (Chapter 13).
[43] Batet M, Valls A, Gibert K. Improving classical clustering with ontologies. In: Proceedings of the 4th World conference of the IASC. Japan; 2008. p. 137–46.