Ontology-based semantic similarity: a new feature-based approach

David Sánchez[1], Montserrat Batet, David Isern, Aida Valls

Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA). Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26. 43007 Tarragona (Spain)

Abstract. Estimation of the semantic likeness between words is of great importance in many applications dealing with textual data such as natural language processing, knowledge acquisition and information retrieval. Semantic similarity measures exploit knowledge sources as the base to perform the estimations. In recent years, ontologies have grown in interest thanks to global initiatives such as the Semantic Web, offering an structured knowledge representation. Thanks to the possibilities that ontologies enable regarding semantic interpretation of terms many ontology-based similarity measures have been developed. According to the principle in which those measures base the similarity assessment and the way in which ontologies are exploited or complemented with other sources several families of measures can be identified. In this paper, we survey and classify most of the ontology-based approaches developed in order to evaluate their advantages and limitations and compare their expected performance both from theoretical and practical points of view. We also present a new ontology-based measure relying on the exploitation of taxonomical features. The evaluation and comparison of our approach's results against those reported by related works under a common framework suggest that our measure provides a high accuracy without some of the limitations observed in other works.

Keywords: semantic similarity, semantic relatedness, ontologies, feature-based similarity, WordNet

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

1. Introduction

With the enormous success of the Information Society and the World Wide Web, the amount of textual electronic information available has significantly increased. As a result, computer understanding of text has acquired great interest in the research community in order to enable a proper exploitation, management, classification or retrieval of textual data.

One of the most basic problems when aiming to interpret textual data is the assessment of semantic likeness between terms because, as it has been demonstrated in psychological experiments (Goldstone, 1994), it acts as a fundamental organizing principle by which humans organize and classify objects. It is important to note that two different paradigms can be found in the literature. On one hand, *semantic similarity* states how taxonomically near two terms are, because they share some aspects of their meaning (*e.g.*, *dogs* and *cats* are similar to the extend they are mammals). On the other hand, the more general concept of *semantic relatedness* does not necessary rely on a taxonomic relation (*e.g.*, *car* and *wheel* or *pencil* and *paper*); other non taxonomic relationships (*e.g.*, meronymy, antonymy, functionality, cause-effect, etc.) are also considered.

Semantic similarity/relatedness computation has many direct and relevant applications. Some basic natural language processing tasks such as word sense disambiguation (Patwardhan, Banerjee, & Pedersen, 2003), synonym detection (Lin, 1998) or automatic spelling error detection and correction (Budanitsky & Hirst, 2001) rely on the assessment of words' semantic resemblance. Direct applications can be found in the knowledge management field, such as thesauri generation (Curran, 2002), information extraction (M. Y. Chen, Chu, & Chen, 2010; P. Chen, Lin, & Chu, 2011; D. Sánchez & Isern, 2009; Stevenson & Greenwood, 2005) or ontology learning (D. Sánchez, 2010; David Sánchez & Antonio Moreno, 2008; D. Sánchez & A. Moreno, 2008), in which new terms related to already existing concepts, should be acquired from textual resources. The Semantic Web is an especially relevant application area, when dealing with automatic annotation of documents (Cimiano, Handschuh, & Staab, 2004; Chu, Chen, & Chen, 2009; D. Sánchez, Isern, & Millan, 2010) and text clustering (Song, Li, & Park, 2009).

Despite its usefulness, robust measurement of semantic similarity/relatedness between textual terms remains a challenging task (Bollegala, Matsuo, & Ishizuka, 2007). Many works have been developed in the last years, especially with the increasing interest on the Semantic Web and the popularization of ontologies (Lanzenberger, et al., 2008). Proposed methods aim for automatically assessing a numerical score between a pair of terms according to the semantic evidence observed in one or several knowledge sources, which are used as semantic background. Ontologies have been of great interest for the semantic similarity research community as they offer a structured and unambiguous representation of knowledge in the form

of conceptualizations interconnected by means of semantic pointers. These structures can be exploited in order to assess the degree of semantic proximity between terms. According to the principle in which the similarity/relatedness computation is based and the way in which the ontology is exploited and/or complemented with other sources (*e.g.*, thesaurus, domain corpora, etc.), different families of methods can be identified.

In this paper we survey and compare most of the ontology-based similarity/relatedness measures developed in recent years. We tried to collect as much relevant approaches as possible in order to offer an updated and detailed review and comparison of the expected performance of these measures both from theoretical and practical points of view. Concretely, for each family of functions, we identify their main advantages and limitations under the dimensions of expected accuracy, computational complexity, dependency on knowledge sources (type, size, structure-dependency and pre-processing) and parameter tuning.

We also present a new feature-based method for semantic likeness assessment based on the exploitation of the taxonomical knowledge available in an ontology. By considering semantic evidence typically omitted by other feature-based approaches, this measure aims to rival state of the art ontology-based approaches in terms of accuracy, while retaining a low computational complexity. In order to compare all the semantic similarity/relatedness measures in a practical setting and evaluate them against our own approach in an objective manner, we collected the results reported by related works covered in the paper and tested our approach under the same conditions. Several widely used benchmarks have been considered in order to enable an objective comparison.

The rest of the paper is organized as follows. Sections 2 surveys, reviews and compares related works in ontology-based semantic similarity/relatedness estimation, classifying them in different families. Section 3 presents the new method for measuring semantic likeness. Section 4 summarizes the evaluation results reported by related works for the analysed measures and tests the accuracy of our own approach under the same conditions. Section 5 discusses and compares the different measures according to the reported evaluation results. The final section contains the conclusions.

2. Ontology-based semantic similarity/relatedness

Ontologies provide a formal specification of a shared conceptualization (Guarino, 1998). Being machine readable and constructed from the consensus of a community of users or domain experts, they represent a very reliable and structured knowledge source. Due to this reason, and thanks to initiatives such as the Semantic Web, which brought the creation of thousands of domain ontologies (Ding, et al., 2004), ontologies have been extensively exploited in

knowledge-based systems (Valls, Gibert, Sánchez, & Batet, 2010) and, more precisely, to compute semantic likeness.

A paradigmatic example is WordNet (Fellbaum, 1998), a domain-independent and general purpose thesaurus that describes and organizes more than 100,000 general English concepts, which are semantically structured in an ontological fashion. It contains words (nouns, verbs, adjectives and adverbs) that are linked to sets of cognitive synonyms (*synsets*), each expressing a distinct concept (*i.e.,* a word sense). Synsets are linked by means of conceptual-semantic and lexical relations such as synonymy, hypernymy (is-a), six types of meronymy (part-of), antonymy, complementary and so on. The backbone of the network of words is the subsumption hierarchy which accounts more than an 80% of all the modelled semantic links, with a maximum depth of 16 nodes. The result is a network of meaningfully related words, where the graph model can be exploited to interpret the meaning of the concept. In fact, WordNet has been used as the background ontology in all related works.

In this section, we survey and compare related works in ontology-based similarity assessment according to the following classification:

1. Edge-counting approaches.
2. Feature-based measures.
3. Measures based on Information Content.

2.1. Edge counting approaches

Ontologies can be seen as a directed graph in which concepts are interrelated mainly by means of taxonomic (is-a) and, in some cases, non-taxonomic links. By mapping input terms to ontological concepts by means of their textual labels, a straightforward method to calculate their similarity is to compute the minimum *Path Length* connecting their corresponding ontological nodes via is-a links (Rada, Mili, Bichnell, & Blettner, 1989). The longer the path, the more semantically far the terms are.

Let us define $path(a,b)=l_1,....l_k$ as a set of links connecting the terms a and b in a taxonomy. Let $|path(a,b)|=k$ be the length of this path. Then, considering all the possible paths from a to b, their semantic distance as defined by (Rada, et al., 1989) is (1).

$$dis_{rad}(a,b) = \min_{\forall i} |path_i(a,b)| \quad (1)$$

Several variations and improvements of this edge-counting approach have been proposed. On one hand, in addition to this absolute distance between terms, Wu and Palmer (Wu & Palmer, 1994) considered that the relative depth in the taxonomy of the concepts corresponding to the evaluated terms is an important dimension, because concept specializations become less distinct as long as they are recursively refined. So, equally distant pairs of concepts belonging to an upper level of a taxonomy should be considered less similar than those belonging to a lower

lever. Wu and Palmer's measure count the number of is-a links (N_1 and N_2) from each term to their Least Common Subsumer (LCS) (*i.e.,* the most concrete taxonomical ancestor that subsumes both terms) and also the number of is-a links of the LCS to the root (N_3) of the ontology (2).

$$sim_{w\&p}(a,b)\frac{2\times N_3}{N_1+N_2+2\times N_3} \quad (2)$$

Based on the same principle, Leadcock and Chodorow (Leacock & Chodorow, 1998) also proposed a measure that considers both the number of nodes N_p separating the ontological nodes corresponding to terms *a* and *b*, included themselves, and the depth *D* of the taxonomy in which they occur in a non-linear fashion (3).

$$sim_{l\&c}(a,b) = -\log(N_p/2D) \quad (3)$$

Li *et al*., (Li, Bandar, & McLean, 2003) also proposed a similarity measure that combines the shortest path length and the depth of ontology information in a non-linear function (4).

$$sim_{li}(a,b) \ = e^{-\alpha \min_{\forall i}|path_i(a,b)|}.\frac{e^{\beta h}-e^{-\beta h}}{e^{\beta h}+e^{-\beta h}} \quad (4)$$

, where *h* is the minimum depth of LCS in the hierarchy and $\alpha \geq 0$ and $\beta > 0$ are parameters scaling the contribution of shortest path length and depth, respectively. Based on a benchmark data, authors stated that the optimal parameters for the measure with respect to a concrete set of human judgements were: $\alpha = 0.2$; $\beta = 0.6$. However, this is just an empirical finding for a specific setting. It lacks a theoretical basis and cannot be generalized.

Al-Mubaid and Nguyen (H. Al-Mubaid & Nguyen, 2006) proposed a cluster-based measure that combines the minimum *path length* and the *taxonomical depth.* They define clusters for each of the branches in the hierarchy with respect the root node. They measure the common specificity of two terms by subtracting the depth of their LCS from the depth D_c of the cluster (5).

$$CSpec(a,b) = D_c - depth(LCS(a,b)) \quad (5)$$

The common specificity is used to consider that lower level pairs of concept nodes are more similar than higher level pairs, as in Wu and Palmer approach. So, the proposed distance measure (*sem*) is defined as follows (6):

$$dis_{sem}(a,b) = \log((\min_{\forall i} |path_i(a,b)|-1)^{\alpha} \times (CSpec)^{\beta} + k) \quad (6)$$

, where $\alpha > 0$ and $\beta > 0$ are contribution factors of path length and common specify features and *k* is a constant. Authors use $k=1$ because with $k \geq 1$ they proved that the distance is positive. Moreover, in their experiments, they give an equal weight to the contribution of the two components (path length and common specify) by using $\alpha = \beta = 1$.

Both Li *et al*., And Al-Mubaid and Nguyen approaches are often considered in the literature (Petrakis, Varelas, Hliaoutakis, & Raftopoulou, 2006; Pirró, 2009) as "hybrid" approaches, as

they combine several structural characteristics (such as path length, depth and local density) and assign weights to balance the contribution of each component to the final similarity value. Even though their accuracy for a concrete scenario (see evaluation section) is higher than more basic edge-counting measures, they depend on the empirical tuning of weights according to the ontology and input terms.

Hirst and St-Onge (Hirst & St-Onge, 1998) extended the notion of taxonomical edge-counting by considering also non-taxonomic semantic links in the path (*full_path*). All types of relations found in WordNet together with rules that restrict possible semantic chains are considered, along with the intuition that the longer the path and the more changes in relation's direction, the lower the likeness. The following path directions are considered: upward (such as *hypernymy* and *meronymy*), downward (such as *hyponymy* and *holonymy*) and horizontal (such as *antonymy*). The resulting formula is (7)

$$sim_{h\&s}(a,b) = C - full\_path(a,b) - k \times turns(a,b) \quad (7)$$

, where C and k are constants ($C = 8$ and $k = 1$ are used by the authors), and *turns(a, b)* is the number of times the path's direction changes.

Due to the non-taxonomic nature of some of the relations considered during the assessment, Hirst and St-Onge's measure captures a more general sense of *relatedness* than of taxonomical *similarity,* assessed by the approaches detailed above.

The main advantage of the presented measures is their simplicity. They only rely on the graph model of an input ontology whose evaluation requires a low computational cost (in comparison to approaches dealing with text corpora, see Section 2.3).

However, several limitations hamper their performance. First, edge-counting measures only consider the shortest path between concept pairs. When they are applied to wide and detailed ontologies such as WordNet that incorporate multiple taxonomical inheritance, the result is that several taxonomical paths are not taken into account. Other features also influencing the concept semantics, such as the number and distribution of common and non-common taxonomical ancestors are neither considered. As a result, by taking only the minimum path between concepts, many of the taxonomical knowledge explicitly modelled in the ontology is omitted.

Another problem of path-based measures typically admitted (Bollegala, Matsuo, & Ishizuka, 2009; Wan & Angryk, 2007) is that they rely on the notion that all links in the taxonomy represent a uniform distance. In practice, the semantic distance among concept specializations/generalizations in an ontology would depend on the degree of granularity and taxonomic detail implemented by the knowledge engineer.

2.2. Feature-based measures

Feature-based methods try to overcome the limitations of path-based measures regarding the fact that taxonomical links in an ontology do not necessary represent uniform distances. This is addressed by considering the degree of overlapping between sets of ontological features. As a result, they are more general and, potentially, they could be applied in cross ontology similarity estimation settings (*i.e.*, when concept pairs belong to two different ontologies), a situation in which edge-counting methods cannot be directly applied (Petrakis, et al., 2006).

So, on the contrary to edge-counting measures which, as stated above, are based on the notion of minimum path distance, feature-based approaches assess similarity between concepts as a function of their properties. This is based on the Tversky's model of similarity (8), which, derived from the set theory, takes into account common and non common features of compared terms, subtracting the latter from the former ones. In fact, common features tend to increase similarity and non-common ones tend to diminish it (Tversky, 1977). Formally, let *Ψ(a)* and *Ψ(b)* be the features of terms *a* and *b* respectively, let $\Psi(a) \cap \Psi(b)$ be the intersection between those two sets of features, and $\Psi(a) \setminus \Psi(b)$ the set obtained when eliminating the elements of *Ψ(b)* from the set of features of concept *a, Ψ(a)*. Then, the similarity between *a* and *b* is proposed to be computed as a function of *Ψ(a)* ∩ *Ψ(b), Ψ(a)\Ψ (b)* and *Ψ(b)\ Ψ (a)* as.

$$sim_{tve}(a,b) = \alpha \cdot F(\Psi(a) \cap \Psi(b)) - \beta \cdot F(\Psi(a) \setminus \Psi(b)) - \gamma \cdot F(\Psi(b) \setminus \Psi(a)) \quad (8)$$

, where *F* is a function that reflects the salience of a set of features, and *α*, *β* and *γ* are parameters that weight the contribution of each component.

The definition of the set of features is crucial in this model. The existing approaches rely on information that is available in ontologies, in particular the set of synonyms (called *synsets* in Wordnet), definitions (*i.e.*, *glosses*, containing textual descriptions of word senses) and different kinds of semantic relationships are considered.

In Rodriguez and Egenhofer (Rodríguez & Egenhofer, 2003), the similarity is computed as the weighted sum of similarities between synsets, features (*e.g.,* meronyms, attributes, etc.) and neighbour concepts (those linked via semantic pointers) of evaluated terms (9).

$$sim_{r\&e}(a,b) = w \cdot S_{synsets}(a,b) + u \cdot S_{features}(a,b) + v \cdot S_{neighborhoods}(a,b) \quad (9)$$

, where *w*, *u* and *v* weight the contribution of each component, which depend on the characteristics of the ontology and *S* represents the overlapping between the different features, computed as:

$$S(a,b) = \frac{|A \cap B|}{|A \cap B| + \gamma(a,b)|A \setminus B| + (1-\gamma(a,b))|B \setminus A|} \quad (10)$$

, where *A*, *B* are the terms evaluated for concepts corresponding to *a* and *b*, *A\B* is the set of terms in *A* but not in *B* and *B\A* the set of terms in *B* but not in *A*. Finally, *γ(a, b)* is computed as a function of the depth of *a* and *b* in the taxonomy as follows:

$$\gamma(a,b)=\begin{cases}\dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a)\le depth(b)\\ 1-\dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a)>depth(b)\end{cases} \tag{11}$$

In Petrakis *et al*., (Petrakis, et al., 2006) a feature-based function called *X-similarity* relies on the matching between synsets and a concept's glosses extracted from WordNet (*i.e.*, words extracted by parsing term definitions). They consider that two terms are similar if the synsets and glosses of their concepts and those of the concepts in their neighbourhood (following semantic relations) are lexically similar. The similarity function is expressed as follows:

$$sim_{X-Similarity}(a,b)=\begin{cases}1, if\ S_{synsets}(a,b)>0\\ \max\{S_{neighborhoods}(a,b),S_{glosses}(a,b)\}, if\ S_{synsets}(a,b)=0\end{cases} \tag{12}$$

The similarity for the semantic neighbours $S_{neighborhoods}$ is calculated as follows:

$$S_{neighborhoods}(a,b)=\max_{i\in SR}\frac{|A_i\cap B_i|}{|A_i\cup B_i|} \tag{13}$$

, where each different semantic relation type (*i.e.*, *is-a* and *part-of* in WordNet) is computed separately and the maximum (considering all the synsets of all concepts up to the root of each hierarchy) is taken.

Equivalently, the similarity for glosses $S_{glosses}$ and synonyms $S_{synsets}$ are both computed as:

$$S(a,b)=\frac{|A\cap B|}{|A\cup B|} \tag{14}$$

, where A and B denote the set of synsets or glosses for the term a and b.

Feature-based measures exploit more semantic knowledge than edge-counting approaches, evaluating both commonalities and differences of compared concepts. However, by relying on features like glosses or synsets (in addition to taxonomic and non-taxonomic relationships), those measures limit their applicability to ontologies in which this information is available. Only big ontologies/thesauri like WordNet include this kind of information. In fact, an investigation of the structure of existing ontologies via the Swoogle ontology search engine (Ding, et al., 2004) reveals that domain ontologies very occasionally model any semantic feature apart from taxonomical relationships.

Another problem is their dependency on the weighting parameters that balance the contribution of each feature. In all cases, those parameters should be tuned according to the nature of the ontology and even to the evaluated terms. This hampers their applicability as a general purpose solution. Only the definition of Petrakis (Petrakis, et al., 2006) does not depend on weighting parameters, as the maximum similarity provided by each feature alone is taken. Even though this adapts the behaviour of the measure to the characteristics of the ontology, the contribution of other features is omitted if only the maximum value is taken at each time.

2.3. Measures based on Information Content

Also acknowledging some of the limitations of edge-counting approaches, Resnik (Resnik, 1995) proposed to complement the taxonomical structure of an ontology with the information distribution of concepts evaluated in input corpora. He exploited the notion of Information Content (IC), by associating appearance probabilities to each concept in the taxonomy, computed from their occurrences in a given corpus. IC of a term *a* is computed according to the negative log of its probability of occurrence, *p*(*a*) (15). In this manner, infrequent words are considered more informative than common ones.

$$IC(a) = -\log P(a) \tag{15}$$

According to Resnik, semantic similarity depends on the amount of shared information between two terms, a dimension which is represented by their Least Common Subsumer (LCS) in an ontology. Two terms are maximally dissimilar if a LCS does not exist (*i.e.*, in terms of edge-counting, it would be not possible to find a path connecting them). Otherwise, their similarity is computed as the IC of the LCS (16).

$$sim_{res}(a,b) = IC(LCS(a,b)) \tag{16}$$

One of the problems of Resnik's metric is that any pair of terms having the same LCS results in exactly the same semantic similarity. Both Lin (Lin, 1998) and Jiang and Conrath (Jiang & Conrath, 1997) extended Resnik's work by also considering the IC of each of the evaluated terms.

Lin proposed that the similarity between two terms should be measured as the ratio between the amount of information needed to state their commonality and the information needed to fully describe them. As a corollary of this theorem, his measure considers, on one hand, commonality in the same manner as Resnik's approach and, on the other hand, the IC of each concept alone (17).

$$sim_{lin}(a,b) = \frac{2 \times sim_{res}(a,b)}{(IC(a) + IC(b))} \tag{17}$$

The measure proposed by Jiang and Conrath is based on quantifying, in some way, the length of the taxonomical links as the difference between the IC of a concept and its subsumer. When comparing term pairs, they compute their distance by subtracting the sum of the IC of each term alone from the IC of their LCS (18).

$$dis_{j\&c}(a,b) = (IC(a) + IC(b)) - 2 \times sim_{res}(a,b) \tag{18}$$

It is important to note that IC-based measures need, in order to behave properly, that the probability of appearance *p* monotonically increases as one moves up in the taxonomy (*i.e.*, $\forall c_i \mid c_j$ *is hypernym of* c_i => $p(c_i) \leq p(c_j)$). This is achieved by computing *p(a)* as the probability of encountering any *instance* of *a* in the given corpus. In practice, each individual occurrence of

any noun in the corpus is counted as an occurrence of each taxonomic class containing it (19) (Resnik, 1995).

$$p(a) = \frac{\sum_{w \in W(a)} count(w)}{N} \quad (19)$$

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

As a result, an accurate computation of concept probabilities requires a proper disambiguation and annotation of each noun found in the corpus. This process is usually done manually to ensure the correctness of the tagging, hampering the scalability and applicability of this approach with large corpora.

Moreover, if either the taxonomy or the corpus changes, re-computations are needed to be recursively executed for the affected concepts. So, it is necessary to perform a manual and time-consuming analysis of corpora and resulting probabilities would depend on the size and nature of input corpora. Moreover, the background taxonomy must be as complete as possible (*i.e.*, it should include most of the specializations of each concept) in order to provide reliable results. Partial taxonomies with a limited scope may not be suitable for this purpose. All those aspects limit the scalability and applicability of those approaches (David Sánchez, Batet, Valls, & Gibert, 2009).

Considering the limitations of IC-based approaches due to their dependency on corpora, some authors tried to intrinsically derive IC values from an ontology. These works rely on the assumption that the taxonomic structure of ontologies like WordNet is organized in a meaningful way, according to the principle of *cognitive saliency* (Blank, 2003). This states that humans specialise concepts when they need to differentiate them from already existing ones. So, concepts with many hyponyms (*i.e.*, specializations) are more general and provide less information than the concepts at the leaves of the hierarchy. From the Information Theory point of view, they consider that abstract ontological concepts appear more probably in a corpus as they subsume many other ones. In this manner, the probability of appearance of a concept (*i.e.* the IC) is estimated as a function of the number of hyponyms and/or their relative depth in the taxonomy.

Seco *et al*., (Seco, Veale, & Hayes, 2004) and Pirró and Seco (Pirró & Seco, 2008) base IC calculations on the number of hyponyms. Being *hypo(a)* the number of hyponyms of the concept *a* and *max_nodes* the number of hyponyms of the root node, they compute IC of a concept in the following way (20):

$$IC_{seco}(a) = 1 - \frac{\log(hypo(a) + 1)}{\log(max\_nodes)} \quad (20)$$

The denominator ensures that IC values are normalized in the range [0..1].

This approach only considers hyponyms of a given concept in the taxonomy; so, concepts with the same number of hyponyms but different degrees of generality appear to be equally similar. In order to tackle the problem, and in the same manner as for edge-counting measures, Zhou *et al*., (Zhou, Wang, & Gu, 2008) proposed to complement hyponym-based IC computation with the relative depth of each concept in the taxonomy. IC of a concept is computed as (21):

$$IC_{zhou}(a) = k\left(1 - \frac{\log(hypo(a)+1)}{\log(max\_nodes)}\right) + (1-k)\left(\frac{\log(depth(a))}{\log(max\_depth)}\right) \quad (21)$$

In addition to *hypo* and *max_nodes*, which has the same meaning as eq. 20, *depth(a)* corresponds to the depth of the concept *a* in the taxonomy and *max_depth* is the maximum depth of the taxonomy. The factor *k* adjusts the weight of the two features involved in the IC assessment. They use k=0.5.

Both ways of intrinsically computing IC have been applied directly on the similarity functions proposed by Resnik, Lin and Jiang and Conrath. Those approaches overcome most of the problems observed for corpus-based IC approaches (specifically, the need of corpus processing and their high data-sparseness) competing and even improving them in terms of accuracy (as it will be stated in the evaluation) when applied over WordNet. However, they require big, and fine grained taxonomies/ontologies with a detailed taxonomical structure in order to properly differentiate concept's IC. For small or very specialized ontologies with a limited taxonomical depth and low branching factor, resulting IC values between concepts would become too homogenous to enable a proper differentiation.

3. A new feature-based measure exploiting taxonomical knowledge

From the study of ontology-based semantic similarity measures presented in the previous section, the basic conclusions that we want to stress are the following:

- Pure ontology-based measures, like the ones based on edge-counting, features and intrinsic IC computation are characterized by their simplicity and computational efficiency as they only exploit the semantic network provided by the ontology. Edge-counting measures are the simplest ones and, in consequence, their accuracies have been surpassed (as it will be stated in the evaluation section) by more complex approaches exploiting additional semantic evidence. Feature-based approaches, however, rely on features which are hardly found in domain ontologies, such as non taxonomic relationships, attributes, synonym sets or glosses. In consequence, their applicability and accuracy depend on the availability of this information.

- Information Content approaches based on semantically-annotated textual data aim to improve pure ontology-based ones by capturing implicit semantic as a function of concept distribution in corpora. However, the association between the words found in a corpus and concepts which are needed to compute accurate concept appearance frequencies is not straightforward, requiring a process of manual word sense tagging for disambiguation. This hampers the applicability of these methods in practise, which are also affected by corpora availability and data sparseness.

In this section we present a new measure that relies on taxonomical features extracted from an ontology. Being a feature-based method, our proposal follows a similar principle as proposed in the Tversky's model (eq. 8), which considers that the similarity between two concepts can be computed as a function their common and differential features.

Differently from previous feature-based approaches presented in section 2.2, we only rely on taxonomic information. This is due to the fact that available ontologies rarely model other kind of knowledge a part from taxonomical relationships (Ding, et al., 2004). As a matter of fact, as stated in section 2, taxonomical knowledge represent more than the 80% of the total amount of relationships modelled in WordNet.

Moreover, on the contrary to other feature-based measures, as only one type of feature will be considered (*i.e.*, taxonomic relationships) no tuning parameters will be used to weight the contribution of potentially scarce semantic features, overcoming one of the limitations observed in related works and improving the generality of our measure.

3.1. Evaluating concept dissimilarity

Our measure considers as features the taxonomical categorization of concepts given by the ontology in order to evaluate the amount of dissimilarity (*i.e.* semantic distance) between concepts. Concretely, we consider that a term can be semantically distinguished from other ones by comparing the set of concepts that subsume it. Using the same notation introduced by Tversky in (Tversky, 1977), the following definitions formalize this idea.

Definition 1. Let C be the set of concepts of an ontology, we define concept subsumption ($\leq$) as a binary relation $\leq : C \times C$. Having two concepts c_i and c_j, $c_i \leq c_j$ is fulfilled if c_i is a hierarchical specialization of c_j or if $c_i = c_j$ (*i.e.*, they are the same concept even though they could be expressed by means of equivalent synonyms).

The fact of including the concept itself in the subsumption relation assumes the notion of dominance as a reflexive relation (Partee, ter Meulen, & Wall, 1990).

Definition 2. The set of taxonomical features describing the concept *a* is defined in terms of the relation $\leq$ as:

$$\phi(a) = \{c \in C \mid a \leq c\} \tag{22}$$

It is important to note that several immediate generalizations (*i.e*. categorizations) per concept may be available in ontologies modelling multiple inheritance, such in WordNet (Devitt & Vogel, 2004) or in detailed domain ontologies such as MeSH or SNOMED-CT (H. Al-Mubaid & Nguyen, 2006). So, in our case, the set of taxonomical features associated to the concept includes *all* the upper categories found when recursively going through all the upper taxonomical paths modelled in the ontology for that concept. Oppositely, ontology-based related works (Jiang & Conrath, 1997; Leacock & Chodorow, 1998; Lin, 1998; Rada, et al., 1989; Resnik, 1995; Wu & Palmer, 1994) deal with the case of multiple taxonomical generalizations per concept by taking the one that defines the maximum similarity with respect to another concept (*e.g*., taking the minimum path between both concepts). In our opinion, this simplification omits a large amount of explicitly available knowledge (*i.e.,* other generalizations and their corresponding taxonomical paths). So, an important characteristic of our definition of taxonomical features (22) is that it does not introduce this limitation.

Given two concepts which are semantically described according to their taxonomic features (those of eq. 22), we consider the degree of disjunction between their feature sets (non-common taxonomical features) as a function of their distance (or dissimilarity) whereas the degree of overlap (common ones) is proportional to their similarity.

Considering the set of differential taxonomical features of *a* with respect to *b* as: $\phi(a) \setminus \phi(b) = \phi(a) - \phi(b) = \{c \in C \mid c \in \phi(a) \wedge c \notin \phi(b)\}$, we formally define the dissimilarity between *a* and *b* as:

Definition 3. The dissimilarity $dis: C \times C \rightarrow \Re$ between *a* and *b* is given by the cardinality of the set of differential features of *a* with respect to *b* and the set of differential features of *b* with respect to *a*:

$$dis(a,b) = \left|\phi(a) \setminus \phi(b)\right| + \left|\phi(b) \setminus \phi(a)\right| \tag{23}$$

In order to enable accurate comparisons between the dissimilarity computed for pairs of concepts corresponding to taxonomical branches with different degrees of taxonomical detail or multiple inheritance, the value given by (23) should be normalized taking into account the total size of the set of features. To include this normalizing factor, we divide the absolute *dis* value by the whole amount of features extracted for both terms. This value corresponds to the sum of

cardinalities of differential and common taxonomical feature sets, that is $|\phi(a) \setminus \phi(b)| + |\phi(b) \setminus \phi(a)| + |\phi(a) \cap \phi(b)|$. This will give a relative dissimilarity value in the [0..1] interval.

As a final consideration, many authors argued that an information theoretic approach, in which semantic features are evaluated in a non linear fashion (H. Al-Mubaid & Nguyen, 2006; Leacock & Chodorow, 1998; Li, et al., 2003), approximates better the concept of similarity. Following a similar principle as intrinsic IC approaches (Seco, et al., 2004; Zhou, et al., 2008), in which the IC of the concept is computed as a non-linear function of the amount of semantic features, we introduce the logarithm in our calculation, obtaining the following formulation.

Definition 4. The normalized dissimilarity between *a* and *b* according to their taxonomical features is calculated as:

$$dis_{norm}(a,b) = \log_2\left(1 + \frac{|\phi(a) \setminus \phi(b)| + |\phi(b) \setminus \phi(a)|}{|\phi(a) \setminus \phi(b)| + |\phi(b) \setminus \phi(a)| + |\phi(a) \cap \phi(b)|}\right) \quad (24)$$

In this expression, the logarithm is calculated by adding 1 to the ratio. In this manner, we avoid infinite values for equivalent terms (*i.e.*, they are equal terms or exact synonyms corresponding to the same concept), which will lead to a zero numerator because there are not unique features differentiating them. Moreover, this expression maintains the value range of the results in the interval [0..1], being the boundary values:
$min\_dis=log_2(1+0)=0,$ for equivalent terms (*i.e.*, $\phi(a) = \phi(b)$ *iff* $(\forall x \mid: x \in \phi(a) \Leftrightarrow x \in \phi(b))$).
$max\_dis=log_2(1+1)=1,$ for maximally different terms with disjoint feature sets resulting in equal numerator and denominator (*i.e.*, $\phi(a) \neq \phi(b)$ *iff* $(\forall x \mid: x \in \phi(a) \Rightarrow x \notin \phi(b))$).

3.2. Example

To illustrate the behavior of our approach, let us consider the following portion of an ontology (Figure 1). At the lowest level we have four concepts: *surfing*, *sailing*, *swimming* and *sunbathing*. All of them are activities of leisure in the beach, but some of them are also sports related to wind and/or water. If only the minimal path connecting the concepts is considered, all of them are at the same distance because they are all brothers with respect to the concept *beach leisure*. So, this will be the answer given by the path length measures presented in section 2.1. With our proposal, which takes into account all the subsumers of the concepts, we are able to distinguish different levels of dissimilarity, which better captures the semantic distance that one would attach to those terms.

<<Figure 1>>

Figure 1. Ontology example.

In our case, the set of features generated for the following four concepts are:

$\phi(surfing) = \{surfing, wind, water, sport, activity, beach\_leisure\}$
$\phi(sailing) = \{sailing, wind, water, sport, activity, beach\_leisure\}$
$\phi(swimming) = \{swimming, water, sport, activity, beach\_leisure\}$
$\phi(sunbathing) = \{sunbathing, beach\_leisure, activity\}$

From these sets of features we can calculate the dissimilarity between any of those pairs of terms. Table 1 contains the results of applying eq. 24. As the measure is symmetric, we present only the lower triangular distance matrix. To give an example, the dissimilarity between *sailing* and *sunbathing* is calculated as follows:

$$dis_{norm}(sailing, sunbathing) = \log_2\left(1 + \frac{4+1}{4+1+2}\right) = 0.78$$

, where the number of differential taxonomical features of *sailing* with respect to *sunbathing* is 4 (*i.e.*, *sailing, wind, water* and *sport*) and the number of differential features of *sunbathing* with respect to *sailing* is 1 (*i.e.*, *sunbathing*), and the set of common elements has a cardinality of 2 (*i.e.*, *beach_leisure* and *activity*).

Table 1. Dissimilarities calculated with dis_{norm} for leaf concepts presented in Figure 1.

Concepts	*Surfing*	*Sailing*	*Swimming*	*Sunbathing*
Surfing	0			
Sailing	0.36	0		
Swimming	0.51	0.51	0	
Sunbathing	0.78	0.78	0.65	0

For the ontology fragment in Figure 1, the concepts *sailing* and *surfing* are the most similar ones (least distance) as it was expected, because they are sports made in the water with the help of the wind, as well as, beach leisure activities (as it is explicitly stated in the ontology via taxonomical relationships). *Swimming* is more similar to *sailing* and *surfing* than *sunbathing* because it is also a water sport. In fact, *sunbathing* is only a relaxing activity in the beach not related with sports, so it is considered the semantically farthest to the rest of concepts. All those semantic evidences are properly captured by our approach by exploiting taxonomical knowledge (especially in the case of multiple inheritance) as it is presented in Table 1.

3.3. Properties

In order to show the validity of the presented measure, we have studied the properties that a dissimilarity measure must fulfill. It is important to note that the fulfilment of those properties is

a requirement if the measure is used in conjunction with some reasoning or data mining techniques (for example, similarity is a core element in case-based reasoning stages: case base building, case retrieval, and even case adaptation (O'Sullivan, Smyth, & Wilson, 2005)). In fact, the coherency of the results obtained in clustering algorithms relying on similarity measures may depend on the fulfilment of those properties (Everitt, Landau, & Leese, 2001). In this respect, several of the measures proposed in related works (especially weighted and feature-based ones (Li, et al., 2003; Rodríguez & Egenhofer, 2003; Tversky, 1977)), do not accomplish basic properties such as minimality, hampering their applicability in some data mining algorithms.

Proposition 1. The function dis_{norm} fulfills the properties of dissimilarity measures (Euzenat & Shvaiko, 2007):

$$\forall a,b \in \mathrm{O}, dis(a,b) \geq 0 \qquad \text{(positiveness)}$$

$$\forall a \in \mathrm{O}, dis(a,a) = 0 \qquad \text{(minimality)}$$

$$\forall a,b \in \mathrm{O}, dis(a,b) = dis(b,a) \qquad \text{(symmetry)}$$

Proof. By adding 1 to the inner expression of the logarithm, values positively range from 0 to 1 so, positiveness is ensured. Equivalent terms will result in an empty set of unique features and in a dis_{norm} value of $log_2(1)=0$, accomplishing the second property that says that the minimum dissimilarity must be given by the comparison of one item with itself. Finally, as *a* and *b* are evaluated according to their feature sets (*i.e.*, the order of the elements in the sets is irrelevant) and all the operations performed over those sets are commutative (*i.e.*, difference and intersection), the measure accomplishes the symmetry property. □

3.4. Dealing with polysemic terms

When input terms are raw words extracted from text, some of them may be polysemous. For general ontologies such as WordNet, polysemic words correspond to several concepts (*i.e.*, one per word sense) which can be found by mapping words to concept synsets. As a matter of fact, in WordNet 2, polysemic nouns correspond to an average of 2.77 concepts[2]. A proper disambiguation of input terms may solve the ambiguity, assigning input words to unique ontological concepts. However, as stated in the introduction, semantic similarity is promoted exactly for applications dealing with various levels of ambiguity because in texts we find words rather than concepts (Budanitsky & Hirst, 2006).

In previous works, polysemic words have been tackled by retrieving all possible concepts corresponding to a term and then, computing individual similarities for each possible pair of

[2] http://wordnet.princeton.edu/wordnet/man2.1/wnstats.7WN.html

concepts and selecting, as the final result, the maximum similarity value obtained. The rationale for this criterion is that in order to evaluate the similarity between two non-disambiguated words (*i.e.*, no context is available), human subjects would pay more attention to their similarities (*i.e.*, most related senses) rather than their differences, as it has been demonstrated in psychological studies (Tversky, 1977). Therefore, we have taken the same approach to solve this problem, taking the minimum dissimilarity value obtained for all the possible combinations.

Definition 5. The generalized dissimilarity measure which is able to deal with polysemic terms is defined as:

$$dis_{generalized}(a,b) = \min_{\substack{\forall a' \in A \\ \forall b' \in B}} dis_{norm}(a',b') \tag{25}$$

, where A is the set of concepts (*i.e.*, word senses) for the term a, and equally for term b.

Proposition 2. The function $dis_{generalized}$ fulfills the properties of dissimilarity measures: positiveness, minimality and symmetry.

Proof. The proof is straightforward considering that dis_{norm} fulfills the positiveness, minimality and symmetry as the minimum operator keeps the properties when applied to it. □

4. Results

As stated in (Bollegala, et al., 2009), an objective evaluation of the accuracy of a semantic similarity function is difficult because the notion of similarity is a subjective human judgement. In order to enable fair comparisons, several authors created evaluation benchmarks consisting of word pairs whose similarity were assessed by a set of humans. Rubenstein and Goodenough (Rubenstein & Goodenough, 1965) defined the first experiment in 1965 in which a group of 51 students, all native English speakers, assessed the similarity of 65 word pairs selected from ordinary English nouns on a scale from 0 (semantically unrelated) to 4 (highly synonymous). Miller and Charles (Miller & Charles, 1991) re-created the experiment in 1991 by taking a subset of 30 noun pairs which similarity was reassessed by 38 undergraduate students. The correlation obtained with respect to Rubenstein and Goodenough experiment was 0.97. Resnik (Resnik, 1995) replicated again the same experiment in 1995, in this case, requesting 10 computer science graduate students and post-doc researchers to assess similarity. The correlation with respect to Miller and Charles results was 0.96. Finally, Pirro (Pirró, 2009) replicated and compared the three above experiments in 2009, involving 101 human subjects, both English and non-English native speakers. He obtained an average correlation of 0.97 with respect to Rubenstein and Goodenough experiment, and 0.95 with respect to Miller and Charles

experiment. It is interesting to see the high correlation obtained between the experiments even though being performed in a period of more than 40 years and with heterogeneous sets of people. This states that similarity between selected words is stable over the years, making them a reliable source for comparing measures.

In fact, Rubenstein and Goodenough and Miller and Charles benchmarks have become *de facto* standard tests to evaluate and compare the accuracy of similarity measures. Authors quantify the accuracy of their measures by computing the correlation between the similarity ratings reported in these benchmarks against those obtained by means of the computerized assessments. If the two ratings are exactly the same which means that the similarity function perfectly mimics human judgements, correlation coefficient will be 1, whereas 0 means that automatic assessments are unrelated to human opinions. Spearman's and Pearson's correlations coefficients have been commonly used in the literature; both are equivalent if ratings sets are ordered (which is the case). They are also invariant to linear transformations which may be performed over results such as a change between distance and similarity by changing the sign of the value or normalizing values in a range. The use of these benchmarks and the correlation coefficient as a measure of evaluation enables an objective comparison between measures.

In order to evaluate the accuracy of related works commented in section 2, we have taken the correlation values originally reported by related works for Rubenstein and Goodenough and Miller and Charles benchmarks (when available) and summarized them in Table 2. In the case in which a concrete measure depends on certain parameters (such as weights or corpora selection/processing) the best correlation value reported in authors' experiments according to optimum parameter tuning was compiled. It is important to note that, even though some of them rely on different knowledge sources (such as tagged corpora), all measures use WordNet as background ontology. Concretely, WordNet 2 is the most common version used in related works. In cases in which original authors used an older version (WordNet 2 was released in July 2003), we took a more recent replication of the measure evaluation performed by another author in order to enable a fair comparison. At the end, we picked up correlation results reported by authors in papers published from 2004 to 2009.

In order to evaluate and compare our approach against related works under the same conditions, we applied it to the two benchmarks also using WordNet 2 as ontology. The correlation values obtained in our case are shown in the last row of Table 2.

Table 2. Correlation values for each measure. From left to right: authors, measure type, correlation for Miller and Charles benchmark, correlation for Rubenstein and Goodenough benchmark and reference in which those correlations where reported

Measure	**Type**	**M&C**	**R&G**	**Evaluated in**
Rada *et al*., (*path length*)	Edge-counting	0.59	N/A	(Petrakis, et al., 2006)
Wu and Palmer	Edge-counting	0.74	N/A	(Petrakis, et al., 2006)
Leacock and Chodorow	Edge-counting	0.74	0.77	(Patwardhan & Pedersen, 2006)
Li *et al*.,	Edge-counting	0.82	N/A	(Petrakis, et al., 2006)
Al-Mubaid and Nguyen (*sem*)	Edge-counting	N/A	0.815	(Hisham Al-Mubaid & Nguyen, 2009)
Hirst and St-Onge	Edge-counting	0.78	0.81	(Wan & Angryk, 2007)
Rodriguez and Egenhofer	Feature	0.71	N/A	(Petrakis, et al., 2006)
Tversky	Feature	0.73	N/A	(Petrakis, et al., 2006)
Petrakis *et al*., (*X-similarity*)	Feature	0.74	N/A	(Petrakis, et al., 2006)
Resnik	IC (corpus)	0.72	0.72	(Patwardhan & Pedersen, 2006)
Lin	IC (corpus)	0.7	0.72	(Patwardhan & Pedersen, 2006)
Jiang and Conrath	IC (corpus)	0.73	0.75	(Patwardhan & Pedersen, 2006)
Resnik (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.829	(Zhou, et al., 2008)
Lin (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.845	(Zhou, et al., 2008)
Jiang and Conrath (IC computed as Seco *et al*.,)	IC (intrinsic)	N/A	0.823	(Zhou, et al., 2008)
Resnik (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.842	(Zhou, et al., 2008)
Lin (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.866	(Zhou, et al., 2008)
Jiang and Conrath (IC computed as Zhou *et al*.,)	IC (intrinsic)	N/A	0.858	(Zhou, et al., 2008)
Our approach (eq. 25)	**Feature**	**0.83**	**0.857**	**-**

5. Discussion

From the results reported in Table 2, in the following, we will make a comparative analysis of the different measures covered in this paper. Together with their accuracy, their main advantages and drawbacks from the application point of view will be discussed.

The basic path length measure (Rada, et al., 1989) presents the lowest accuracy (0.59) due to the fact that absolute lengths of the paths connecting two concepts may not accurately represent their specificity. This is the case of WordNet, since concepts higher in the hierarchy are more general than those lower in the hierarchy (Pirró, 2009). As a result, other edge-counting approaches also exploiting the relative depth of the taxonomy (Wu and Palmer (Wu & Palmer, 1994), Leadcock and Chodorow (Leacock & Chodorow, 1998)) offer a higher accuracy (0.74). It is remarkable the correlation values obtained by Li (Li, et al., 2003) and Al-Mubaid and

Nguyen approaches (H. Al-Mubaid & Nguyen, 2006), which combine the length of the path with the depth of the concepts in a weighted and non-linear manner. However, they rely on empirical parameters whose values have been experimentally determined to optimize the accuracy for the evaluated benchmark, hampering their generality. Hirst and St-Onge (Hirst & St-Onge, 1998) presents a similar behaviour, also relying on tuning parameters but, in this case, using non-taxonomic relationships that consider a more general concept of relatedness.

Feature-based methods try to overcome the limitations of path-based measures by considering different kinds of ontological features. The problem, which has been also noted for some edge-counting measures, is their dependence on the parameters introduced to weight the contribution of each feature (for Rodriguez and Egenhofer (Rodríguez & Egenhofer, 2003) and Tversky (Tversky, 1977) approaches). Correlation values are, however, very similar to those offered by edge-counting measures (0.71-0.74) in these benchmarks. This can be motivated by the fact that they rely on concept features, such as synsets, glosses or non-taxonomic relationships which have secondary importance in ontologies like WordNet in comparison with taxonomical knowledge. In fact, those kind of features are scarce in ontologies (Ding, et al., 2004), which causes those approaches are based on partially modelled knowledge. As a result, those measures, even being more complex, are not able to significantly outperform the state of the art of edge-counting measures.

For IC-based measures, we observe that approaches relying on an intrinsic computation of IC (based on the number of concept hyponyms) clearly outperform approaches relying on corpora (0.72 vs. 0.84, in average). This is very convenient as corpora dependency seriously hampers the applicability of classic IC measures. The difference between both ways of computing IC is caused by two factors. Firstly, the data sparseness problem that appear when relying on tagged corpora (which would be necessary small due to manual tagging) to obtain accurate concept appearance frequencies. Secondly, the fact that WordNet's taxonomy is detailed and fine-grained, which enables an accurate estimation of a term's generality as a function of its number of hyponyms. With regards to the performance of each measure, Lin (Lin, 1998) tends to improve Resnik (Resnik, 1995) one when IC is computed intrinsically, as the former is able to differentiate terms with identical LCS but different taxonomical depths. With regards to the way in which intrinsic IC is computed, more complex approaches also exploiting relative depth and relying on weighting parameters (Zhou *et al*., (Zhou, et al., 2008)) offer the highest accuracy (0.86).

Comparing the proposed measure with other ontology-based works, one can note that our approach's accuracy surpasses the basic edge-counting approaches (0.83 vs. 0.74). In general, in complex and detailed ontologies like WordNet, where multiple taxonomical paths can be found connecting concept pairs (overlapping hierarchies), path-based measures waste explicitly available taxonomical knowledge as only the minimum path is considered as an indication of

distance. Only the Li *et al*.,'s measure is able to achieve a very similar accuracy when the appropriate scaling parameters are empirically chosen.

Feature-based approaches' correlations are also surpassed (0.84 vs. 0.74), even though they are based on other non-taxonomical features and weighting parameters. This shows that taxonomical knowledge plays a more relevant role in stating term similarity than other more scarce features which are typically poorly considered in available ontologies.

The same situation is repeated for corpus-based IC measures (0.84 vs. 0.73) showing that the exploitation of high quality taxonomical knowledge available in ontologies provides even more reliable semantic evidences than unstructured textual resources. This is coherent to what is observed for approaches computing IC in an intrinsic manner, which, conceptually, follow a similar principle as our approach. In their case, similarity is computed as a function on the number of hyponyms whereas in our case it is estimated as a function of overlapping and non-overlapping hypernyms. Moreover, Lin and Jiang and Conrath measures computing IC intrinsically follow the same principle as feature-based measures: similarity is proportional to feature overlapping (in their case, represented by the IC of the LCS) and inversely proportional to the differences (in their case, the IC of each individual term). So, if the IC is computed from taxonomical evidences (*i.e.*, number of hyponyms) it is coherent that their correlation values are similar as those of our approach. The only case in which they surpass our measure's correlation is when IC is computed as Zhou *et al.*'s (Zhou, et al., 2008), in which a weighting parameter is introduced to optimize the assessment.

Summarizing, our measure is able to provide a high accuracy without any dependency on data availability, data pre-processing or tuning parameters for a concrete scenario. As it only relies on the most commonly available ontological feature, our measure ensures its generality as a domain-independent proposal. At the same time, it retains the low computation complexity and lack of constraints of edge-counting measures as it only requires retrieving, comparing and counting ontological subsumers. This ensures its scalability when it must be used in engineering or data mining applications, which may require dealing with large sets of terms (Armengol, 2009; Batet, Valls, & Gibert, 2008).

Compared to other approaches based on taxonomical knowledge, the exploitation of the whole amount of unique and shared subsumers seems to give solid semantic evidences of semantic resemblance. First, the distinctive features implicitly include information about the different paths connecting the pair of terms. In the same manner, the depth of the Least Common Subsumers of those concepts is implicitly included in the set of shared subsumers (*i.e.*, the deeper the LCS, the higher the amount of common features). Other features that have been identified in the literature, such as relative taxonomical densities and branching factors, are also implicitly considered, being all of them useful dimensions to assess semantic similarity.

As any other ontology-based measure, the final accuracy will depend on the coverage, 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 inheritance, the accuracy improvements of our approach with respect to measures based on the minimum path are likely to be less noticeable. Fortunately, large and broad ontologies are being developed, like WordNet as a general purpose description of concepts, or the UMLS repository in the medical context.

6. Conclusions

As it has been explained in the introduction, semantic similarity assessment is a crucial component embedded in many applications framed in the artificial intelligence research area. This paper provides an up-to-date survey of ontology-based semantic similarity measures that can be used to estimate the resemblance between terms.

A new measure based on taxonomical features has been also presented and compared in the context of the survey. In this measure the set of features is built from the categorization (*i.e.*, subsumers) of the concepts modelled in the ontology. In our case, we consider subsumers as labels that describe the meaning of the concept in different levels of generality.

Differently to the other feature-based approaches, this measure only relies on taxonomic ontological knowledge (which is the most commonly available one), lacking of corpora-dependency or parameter-tuning. It is computational efficient as only taxonomical branches are explored and it fulfils the mathematical properties required in many applications for coherent similarity computation (Everitt, et al., 2001; O'Sullivan, et al., 2005).

The paper has analysed, under a common framework, the pros and cons of both related works and our proposal, with the aim of giving some insights on their accuracy, applicability, dependencies and limitations. In addition, a complete comparison of all these measures in a practical setting is reported, using the two widely used benchmarks. The conclusions extracted from those analyses would help practitioners in selecting the measure that better fits with the requirements of a concrete application. In particular, the results reported by our measure for the two benchmarks suggest a promising accuracy, improving the correlations reported by most of other ontology-based approaches, while minimizing the constraints that may hamper its applicability both from the computational efficiency and resource-dependency points of view. For this reason, as future work, we want to study how the inclusion of this measure can improve the results in some concrete applications. In particular, we are studying semantically grounded

data mining processes and statistical disclosure control methods, which may directly benefit from a more accurate similarity assessment.

Acknowledgements

This work has been partially supported by the Universitat Rovira i Virgili (a pre-doctoral grant of M. Batet, a post-doctoral grant of D. Isern and 2009AIRE-04) and 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).

References

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

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

Armengol, E. (2009). Using explanations for determining carcinogenecity in chemical compounds. *Engineering Applications of Artificial Intelligence, 22*, 10-17.

Batet, M., Valls, A., & Gibert, K. (2008). Improving classical clustering with ontologies. In *4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis, IASC 2008* (pp. 137-146). Yokohama, Japan: International Association for Statistical Computing.

Blank, A. (2003). 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* (pp. 37-66). Berlin, Germany: Mouton de Gruyter.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring Semantic Similarity between Words Using Web Search Engines. In C. Williamson & M. E. Zurko (Eds.), *16th international conference on World Wide Web, WWW 2007* (pp. 757-766). Banff, Alberta, Canada ACM.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2009). 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* (pp. 803–812). Singapore, Republic of Singapore: ACL and AFNLP.

Budanitsky, A., & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, Second*

meeting of the North American Chapter of the Association for Computational Linguistics (pp. 10-15). Pittsburgh, USA.

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

Cimiano, P., Handschuh, S., & Staab, S. (2004). Towards the self-annotating web. In *13th international conference on World Wide Web, WWW 2004* (pp. 462 - 471). New York, USA: ACM.

Curran, J. R. (2002). Ensemble Methods for Automatic Thesaurus Extraction. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2002* (pp. 222–229). Philadelphia, PA, USA: Association for Computational Linguistics.

Chen, M. Y., Chu, H. C., & Chen, Y. M. (2010). Developing a semantic-enable information retrieval mechanism *Expert Systems with Applications, 37*, 322-340.

Chen, P., Lin, S. J., & Chu, Y. C. (2011). Using Google latent semantic distance to extract the most relevant information *Expert Systems with Applications, 38*, 7349-7358.

Chu, H. C., Chen, M. Y., & Chen, Y. M. (2009). A semantic-based approach to content abstraction and annotation for content management *Expert Systems with Applications, 36*, 2360-2376.

Devitt, A., & Vogel, C. (2004). The topology of WordNet: Some Metrics. In P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (Eds.), *The Second Global Wordnet Conference, GWC 2004* (pp. 106-111). Brno, Czech Republic: Masaryk University.

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

Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching*: Springer Verlag.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis*. London: Arnold.

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

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 3-28.

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

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305–332): MIT Press.

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

Lanzenberger, M., Sampson, J., Kargl, H., Wimmer, M., Conroy, C., O'Sullivan, D., Lewis, D., Brennan, R., Ramos Gargantilla, J. Á., Gómez-Pérez, A., Fürst, F., Trichet, F., Euzenat, J., Polleres, A., Scharffe, F., & Kotis, K. (2008). Making Ontologies Talk: Knowledge Interoperability in the Semantic Web. *IEEE Intelligent Systems, 23*, 72-85.

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

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

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

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

O'Sullivan, D., Smyth, B., & Wilson, D. C. (2005). Understanding case-based recommendation: A similarity knowledge perspective. *International Journal on Artificial Intelligence Tools, 14*, 215-232.

Partee, B., ter Meulen, A., & Wall, R. (1990). *Mathematical Methods in Linguistics*: Kluwer Academic Publishers.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). 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* (Vol. 2588, pp. 241-257). Mexico City, Mexico: Springer Berlin / Heidelberg.

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together* (pp. 1-8). Trento, Italy.

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

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

Pirró, G., & Seco, N. (2008). Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. In R. Meersman & Z. Tari (Eds.), *OTM 2008 Confederated International Conferences CoopIS, DOA, GADA, IS, and ODBASE 2008* (Vol. 5332, pp. 1271-1288). Monterrey, Mexico: Springer Berlin / Heidelberg.

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

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

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

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

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

Sánchez, D., Batet, M., Valls, A., & Gibert, K. (2009). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems, 35*, 383-413.

Sánchez, D., & Isern, D. (2009). Automatic extraction of acronym definitions from the Web. *Applied Intelligence, doi: 10.1007/s10489-009-0197-4* (in press).

Sánchez, D., Isern, D., & Millan, M. (2010). Content Annotation for the Semantic Web: an Automatic Web-based Approach. *Knowledge and Information Systems, doi:10.1007/s10115-010-0302-3*, (in press).

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

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

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

Song, W., Li, C. H., & Park, S. C. (2009). Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications, 36*, 9095-9104.

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

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

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

Wan, S., & Angryk, R. A. (2007). Measuring Semantic Similarity Using WordNet-based Context Vectors. In M. El-Hawary (Ed.), *IEEE International Conference on Systems, Man and Cybernetics, SMC 2007* (pp. 908 - 913). Montreal, Quebec, Canada: IEEE Computer Society.

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

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