Ontology-driven web-based semantic similarity

David Sánchez · Montserrat Batet ·
Aida Valls · Karina Gibert

Abstract Estimation of the degree of semantic similarity/distance between concepts is a very common problem in research areas such as natural language processing, knowledge acquisition, information retrieval or data mining. In the past, many similarity measures have been proposed, exploiting explicit knowledge—such as the structure of a taxonomy—or implicit knowledge—such as information distribution. In the former case, taxonomies and/or ontologies are used to introduce additional semantics; in the latter case, frequencies of term appearances in a corpus are considered. Classical measures based on those premises suffer from some prob-lems: in the first case, their excessive dependency of the taxonomical/ontological structure; in the second case, the lack of semantics of a pure statistical analysis of occurrences and/or the ambiguity of estimating concept statistical distribution from term appearances. Measures based on Information Content (IC) of taxonomical concepts combine both approaches. However, they heavily depend on a properly pre-tagged and disambiguated corpus according to the ontological entities in order to compute accurate concept appearance probabilities. This limits the applicability of those measures to other ontologies –like specific domain ontologies- and massive corpus –like the Web-. In this paper, several of the presented issues are analyzed. Modifications of classical similarity measures are also proposed. They are based on

D. Sánchez (✉) · M. Batet · A. Valls
Department of Computer Science and Mathematics, Universitat Rovira i Virgili (URV),
Avda. Països Catalans, 26, 43007 Tarragona, Spain
e-mail: david.sanchez@urv.cat

M. Batet
e-mail: montserrat.batet@urv.cat

A. Valls
e-mail: aida.valls@urv.cat

K. Gibert
Department of Statistics and Operations Research, Universitat Politècnica de Catalunya,
Campus Nord, Ed.C5, c/Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: karina.gibert@upc.edu

a contextualized and scalable version of IC computation in the Web by exploiting taxonomical knowledge. The goal is to avoid the measures' dependency on the corpus pre-processing to achieve reliable results and minimize language ambiguity. Our proposals are able to outperform classical approaches when using the Web for estimating concept probabilities.

Keywords Semantic similarity · Ontologies · Information content · Web · Knowledge discovery

1 Introduction

The computation of the semantic similarity/distance between concepts has been a very active trend in computational linguistics. As stated in (Patwardhan and Pedersen 2006), *semantic similarity* gives a clue of the degree of *taxonomical* alikeness between concepts and it can be distinguished from more general *relatedness* approaches which consider other kinds of inter-concept semantic relationships (i.e. non-taxonomic). Similarity computation is an important issue which have many direct applications, such as, word-sense disambiguation (Resnik 1999), document categorization or clustering (Cilibrasi and Vitanyi 2006), word spelling correction (Budanitsky and Hirst 2006), automatic language translation (Cilibrasi and Vitanyi 2006), ontology learning (Sánchez 2008) or information retrieval (Lee et al. 1993). In addition, other knowledge-related fields are interested in the measurement of semantic similarity. For example, ontology-driven data mining (Tadepalli et al. 2004) or privacy preserving through anonymization based on semantic knowledge (Ruch et al. 2000).

In general, the assessment of concept's similarity is based on the estimation of semantic evidence observed in a knowledge source. So, background knowledge is needed in order to measure the degree of similarity between concepts. The more background knowledge is available (i.e. textual corpus, dictionaries, taxonomies/ontologies, etc.) and the more pre-processing of the data (e.g. manual tagging, disambiguation, etc.), the better the estimation will be. However, excessive dependency of a pre-processed data may hamper the generality or applicability of the measure due to the manual knowledge management bottleneck.

1.1 Related work

In the literature, we can distinguish several different approaches to compute concept semantic similarity according to the techniques employed and the knowledge exploited to perform the assessment.

First, there are unsupervised approaches in which semantics are estimated from the information distribution of terms (instead of concepts) in a given corpus (Etzioni et al. 2005; Landauer and Dumais 1997). Statistical analysis and shallow linguistic parsing are used to measure the degree of co-occurrence between terms which is used as an estimation of similarity (Lemaire and Denhière 2006). These are collocation-based measures (Ferreira da Silva and Lopes 1999; Church et al. 1991) following the premise that term co-occurrence is an evidence of their relatedness. These measures need a corpus as general as possible in order to estimate social-scale word usage.

Particularly, the Web has been exploited by these measures (Turney 2001; Downey et al. 2007). However, due to the lack of semantic analysis over the text, problems about language ambiguity (i.e. polysemic terms) or misinterpretation of term co-occurrences compromise the results.

Other trends exploit structured representations of knowledge as the base to compute similarities. Typically, subsumption hierarchies, which are a very common way to structure knowledge (Gómez-Pérez et al. 2004), have been used for that purpose. The evolution of those basic semantic models has given the origin to ontologies in which many types of relationships and logical descriptions can be specified to formalize knowledge (Guarino 1998). Nowadays, with the development of the Semantic Web (Berners-lee et al. 2001), in addition to massive and general purpose linguistic ontologies such as WordNet (Fellbaum 1998), many domain ontologies have been developed and are available through the Web (Ding et al. 2004).

From the similarity point of view, taxonomies and, more generally, ontologies, provide a graph model in which semantic interrelations are modeled as links between concepts. Many approaches have been developed to exploit this geometrical model, computing concept similarity as inter-link distance (Wu and Palmer 1994; Rada et al. 1989; Leacock and Chodorow 1998). *Similarity* measures considered in this paper exploit is-a taxonomical links, whereas more general *relatedness* measures may exploit other types of semantic links (e.g. meronyms). The main problem of those path-length-based approaches is that they heavily depend on the degree of completeness, homogeneity and coverage of the semantic links represented in the ontology. General massive ontologies such as WordNet, with a relatively homogeneous distribution of semantic links and good inter-domain coverage, are the ideal environment to apply those measures (Jiang and Conrath 1997). For other more specific domain ontologies with a limited scope, the graph model may be partial; in this case, path-based measures will be affected by the bias introduced by the partial knowledge modeling (Cimiano 2006). It is worth to note that the presence of a semantic link between two concepts gives an evidence of a relationship but not about their *semantic distance* (i.e. all individual links have the same *length* and, in consequence, represent uniform distances (Bollegala et al. 2007)).

On the other hand, there exist other ontology-based similarity measures which combine the knowledge provided by an ontology and the information distribution of concepts in a corpus to compute their Information Content (IC). IC measures the amount of information provided by a given term from its probability of appearance in a corpus. So, infrequent words are more informative than common ones. Based on this premise, Resnik (1995) presented a seminal work in which the similarity between two terms is estimated as the amount of taxonomical information they share in common. In a taxonomy, this information is represented by the Least Common Subsumer (LCS) of both terms. So, the computation of the IC of the LCS results in an estimation of the similarity of the subsumed terms. The more specific the subsumer is (higher IC), the more similar the subsumed terms are, as they share *more information*. Several variations of this measure have been developed (as presented in Section 2). Using an appropriate corpus as background (such as SemCor (Miller et al. 1993)) those measures outperform path length-based ones (Patwardhan and Pedersen 2006).

However, Resnik-like measures heavily depend on two aspects: (1) the way of computing the IC which, at the same time, depends on the corpus and (2) the coherence between the IC values of ontological concepts and the organization of the subsumption hierarchy. Regarding the first aspect, Resnik-based similarity measures employ a manually pre-tagged corpus associated to WordNet nouns in order to avoid language ambiguity (mainly polysemy and synonymy) (Miller et al. 1993). However, this hampers the applicability of the approach to rare words for which no data is available in the corpus. Regarding the latter aspect, the IC value of the compared concepts should monotonically increase according to their degree of specialization. These measures solve this last issue by recursively adding concept occurrences from all of its subsumed terms.

As a consequence of the presented issues, Resnik-like measures heavily depend on both the ontology (which should be as complete as possible) and the pre-processed corpus data in order to achieve accurate results.

1.2 Our contribution

In this paper, we present modified versions of classical semantic similarity measures (based on taxonomical knowledge) that overcome the presented limitations. On one hand, in order to minimize the corpus dependency and, in consequence, the coverage limitations of Resnik-based measures, we will not rely on pre-processed data. In fact, a completely unprocessed and massive corpus as the Web will be exploited to assess reliable estimations of concept appearance probabilities. On the other hand, unlike unsupervised collocation-based approaches, taxonomical knowledge will be employed to minimize the ambiguity of term co-occurrences in the corpus.

In order to achieve reliable similarity estimations from the Web without manual pre-tagging or explicit disambiguation, our approach proposes a new way of computing concept IC from the Web in a taxonomically coherent manner (i.e. monotonically increasing as concepts are specialized) and minimizing language ambiguity. As it will be shown in the evaluation section, using this strategy, the modified measures are able to outperform classical similarity functions when testing them using a standard benchmark (Resnik 1995), WordNet as ontology and the Web as the corpus.

The rest of the paper is organized as follows. Section 2 analyzes the similarity measures based on IC computation of ontological concepts, showing their limitations when an unprocessed corpus as the Web is used for estimating concept appearance probabilities. Section 3 presents a new way of computing concept IC from the Web and shows its application to Resnik-based measures. Section 4 reviews unsupervised collocation measures for similarity assessment, identifying their limitations due to their lack of semantics. Section 5 applies the proposed IC computation over collocation measures in order to minimize language ambiguity. Section 6 evaluates all the modified measures against the classical versions using a standard benchmark based on human judgments. The last section summarizes the approach, presents the conclusions and outlines some of its applications.

2 Analyzing IC-based ontology-driven similarity measures

Information content (IC) of a concept is the inverse to its probability of occurrence. The IC is taken as the negation of the logarithm of the probability $p(a)$ of encountering a concept c in a given corpus (1). In this way, infrequent words obtain a higher IC than more common ones.

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

As mentioned in the introduction, Resnik (1995) introduced the idea of computing the similarity between a pair of concepts as the IC of the *Least Common Subsumer* (LCS) in a given ontology (2), as an indication of the amount of information that concepts share in common. The more specific the subsummer is (higher IC), the more similar the terms are.

$$sim_{res}(a,b) = IC(LCS(a,b)) \tag{2}$$

The most commonly used extensions to Resnik measure are Lin (1998) and Jiang and Conrath (1997).

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 (3).

$$sim_{lin}(a,b) = \frac{2 \times sim_{res}(a,b)}{(IC(a) + IC(b))} \tag{3}$$

Jiang & Conrath subtract the information content of the LCS from the sum of the information content of the individual concepts (4).

$$dis_{jcn}(a,b) = (IC(a) + IC(b)) - 2 \times sim_{res}(a,b) \tag{4}$$

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

In order to obtain reliable results using those classical approaches, the way in which $p(a)$ is computed is crucial. The presented IC-based measures obtain near baseline results (compared to human judgments (Miller and Charles 1991)) when they estimate word frequencies from SemCor (Miller et al. 1993), a semantically tagged text consisting of 100 passages from the Brown Corpus. Since the tagging scheme was based on WordNet 1.6 (Fellbaum 1998) word sense definition and WordNet is used by those measures to extract the LCS, frequency distribution for each *synset*[1] is very precise. As frequencies of appearances are referred to concepts rather than words, on the one hand, word polysemy does not affect as term appearances are unambiguously associated to WordNet concepts (*synsets*). On the other hand, synonyms are also associated to the appropriate *synset* so they do not negatively affect IC computation. The drawback of using this data is its small size and high data sparseness 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) actually appear in the corpus.[2]

The coherence of the IC computation and the taxonomical structure is the other aspect that should be ensured to maintain the consistency of the similarity computation. Resnik-based measures explicitly introduce the premise that IC of the subsumer

[1] A synset in WordNet groups a set of synonyms and a gloss corresponding to a word sense (i.e. concept).

[2] http://wordnet.princeton.edu/man/wnstats.7WN

must be lower than its specializations. For example, Jiang and Conrath (1997) approximate the probability of co-occurrence of the subsumer and its specialization to the probability of the latter (based on the conceptual inclusion of the taxonomy). So, if the IC associated to each ontological concept does not monotonically increase as concepts are specialized, similarity values would be negatively affected. To guarantee this property, the probability of a concept can be calculated as the sum of the individual occurrences of all the concepts which are subsumed by it (5), as proposed by Resnik (1995).

$$p\left(a\right)=\sum_{n\in specializations\left(a\right)}\frac{count\left(n\right)}{N}, \tag{5}$$

where *specializations(a)* is the set of terms subsumed by concept a and N is the total number of concepts observed in the corpus.

In this manner, subsumers will always be considered as more general—less IC—than their subsumed concepts. This is what we call a *taxonomically coherent IC computation*. However, this forces to recursively compute all the appearances of the subsumed terms before obtaining the IC of the subsumer. If the taxonomy or the corpus change, re-computations of the affected branches are needed, hampering the scalability of the solution when offering up-to-date probabilities. Moreover, the background taxonomy must be as complete as possible so that it includes most of the specializations (e.g. all the possible *mammals*) for a specific concept (e.g. *mammal*) in order to provide reliable results. Partial ontologies with a limited scope may not be suitable for this purpose.

All the mentioned issues show a heavy dependence of the measures from the—limited—corpus used as background and its—even more limited—pre-processing. In addition, they almost force the use of a general purpose and highly detailed ontology as WordNet in order to achieve reliable results.

2.1 Computing IC from a general corpus: the Web

The corpus-dependency of IC-based measures stated above introduces limitations about the applicability of the measures as general purpose similarity assessors. Data sparseness (i.e. the fact that not enough tagged data is available for certain concepts to reflect an appropriate semantic evidence) is the main problem.

Ideally, the robustness of the semantic evidence may be increased by using a bigger and more general corpus like the Web. The Web offers more than 1 trillion of accessible resources which are directly indexed by web search engines[3] (compared to the 100 passages of SemCor (Miller et al. 1993)). It has been demonstrated (Brill 2003) the convenience of using such a wide corpus to improve the sample quality for statistical analysis. Concretely, the amount and heterogeneity of information in the Web are so high that it can statistically approximate the real distribution of information (Cilibrasi and Vitanyi 2006).

The problem is that the analysis of such an enormous repository for computing concept appearances is impracticable. However, the availability of massive Web

[3] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Information Retrieval tools can help in this purpose. The frequency of page counts returned by the search engine divided by the number of indexed pages can be used to estimate the probability of appearance of a term. In fact, Cilibrasi and Vitanyi (2006) claim that the probabilities of web search engine terms approximate the relative frequencies of those searched terms as actually used in society. So, exploiting Web Information Retrieval (IR) tools and concept's usage at a social scale as an indication of its generality, one can estimate the concept probabilities from web hit counts (Turney 2001). Intuitively, the IC of a concept may be estimated from the Web with the ratio presented in Definition 1.

Definition 1 *Web-based Information Content* (*IC_IR*) of a concept '*a*' is defined as:

$$IC\_IR(a) = -\log_2 p_{web}(a) = -\log_2 \frac{hits(a)}{total\_webs}, \tag{6}$$

where $p_{web}(a)$ is the probability of appearance of word '*a*' in a web resource. This probability is estimated from the web hit counts returned by Web IR tool—*hits*—when querying the term '*a*'. *Total_webs* is the total number of resources indexed by a web search engine.

In that case, estimating *concept* probabilities from absolute *term* web hit counts without further—manual—processing can lead to very inaccurate results. Several languages related issues which affect to this estimation can be identified:

1. Absolute word usage in a corpus is a poor estimation of concept probability. This may lead to incoherent concept IC computation with respect to the underlying subsumption hierarchy. For example, as shown in Fig. 1, the word *mammal*, as a

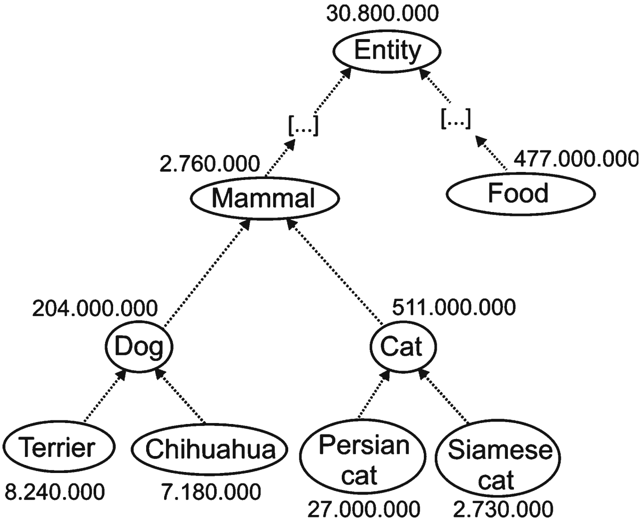

Fig. 1 Portion of an example taxonomy (with *Entity* as the root concept) and occurrence values of concept's terms in the Web computed from Bing hit count (Accessed: Nov. 9th, 2008)

subsumer of *dog* is much less frequent than the later in a general corpus like the Web. This may affect the monotony of the IC_IR associated to the taxonomy. As mentioned in the previous section, this is usually solved in Resnik-based similarity measures by computing all individual occurrences of each concept and adding it to its subsumers. However, implementing this solution for the Web will lead to an enormous amount of web queries to recursively compute occurrences of all the concept's specializations, as well as, a heavy dependence on the corpus and ontology (re-computation will be needed to keep results up-to-date).

2. Language ambiguity may cause different problems: on one hand, different synonyms or lexicalizations of the same concept may result in different IC_IR values (e.g. *dog* is much more frequent than *canis*),[4] introducing bias. On the other hand, the same term may have different senses and, in consequence, correspondences to several concepts. In that last case, the computed term IC_IR will be the sum of IC_IR for all the associated concepts (e.g. IC of *dog* computed from a corpus includes appearances referring to a *mammal* and a *hot dog*, among other possible senses). As mentioned in the previous section, in classical approaches (Resnik 1995; Hotho et al. 2002) those problems were omitted by using a corpus tagged at a concept level based on WordNet *synsets*, rather than a word-level analysis. Therefore, they avoid potentially spurious results when only term (not concept) frequencies are used (Resnik 1995). In a more general approach, where the IC of a concept is computed from estimated term occurrences in the Web, ambiguity may cause inconsistencies if the context is not taken into consideration.

2.2 An example

In order to illustrate the poor estimation of the IC_IR introduced in the previous section, and its effects in the similarity assessment, let us consider the taxonomy presented in Fig. 1 and the estimated term Web appearances obtained from queries performed over the Bing[5] web search engine (more details on the convenience of using this search engine will be provided in the evaluation section).

We have applied the Resnik similarity measure to this taxonomy introducing the IC_IR (Eq. 6), using as concept appearances the Web term hit count presented in Fig. 1. The final concept probabilities are obtained by dividing term appearances by the total amount of indexed resources in the corpus, which in the case of the Web we have considered to be 10^{12} (as introduced in the previous section). The similarity between some pairs of concepts has been estimated. First two dog breeds are compared: *Terrier* and *Chihuahua*. Then, the similarity between *Chihuahua* and *Persian Cat* is calculated. In the latter case, the most specific concept that generalizes

[4] Occurrence of the word *dog* is 204 millions, while *canis* is 2 millions, computed from Bing (Nov. 9th, 2008).

[5] Bing search engine (http://www.bing.com).

them is *Mammal*, according to the ontology given in Fig. 1. The following results have been obtained:

$$\begin{aligned} sim_{res}\_IR\,(terrier, chihuahua) &= IC\_IR\,(LSC\,(terrier, chihuahua)) \\ &= IC\_IR\,(dog) = -\log_2 \frac{hits\,(dog)}{total\_webs} \\ &= -\log_2 \frac{204 \times 10^6}{10^{12}} = 12.25 \\ sim_{res}\_IR\,(chihuahua, persian\_cat) &= IC\_IR\,(LSC\,(chihuahua, persian\_cat)) \\ &= IC\_IR\,(mammal) = -\log_2 \frac{hits\,(mammal)}{total\_webs} \\ &= -\log_2 \frac{2.76 \times 10^6}{10^{12}} = 18.46, \end{aligned}$$

in consequence, we will erroneously conclude that

$$sim_{res}\_IR\,(chihuahua, terrier) < sim_{res}\_IR\,(chihuahua, persian\_cat)\,,$$

because, contrarily to what it is expected in a subsumption hierarchy, *IC_IR(mammal) > IC_IR(dog)*.

The non-monotonic IC_IR values affect even more to measures in which the IC of the LCS is compared against the IC of the evaluated concepts, producing incorrect results (out of range values). Concretely, evaluating the similarity using Lin measure and IC_IR, we obtain the following results:

$$\begin{aligned} sim_{lin}\_IR\,(terrier, chihuahua) &= \frac{2 \times (IC\_IR\,(dog))}{(IC\_IR\,(terrier) + IC\_IR\,(chihuahua))} \\ &= \frac{2 \times \left(-\log_2 \left(\frac{204 \times 10^6}{10^{12}}\right)\right)}{-\log_2 \frac{8.24 \times 10^6}{10^{12}} - \log_2 \frac{7.18 \times 10^6}{10^{12}}} = 0.72 \end{aligned}$$

$$\begin{aligned} sim_{lin}\_IR\,(chihuahua, persian\_cat) &= \frac{2 \times (IC\_IR\,(mammal))}{(IC\_IR\,(chihuahua) + IC\_IR\,(persian\_cat))} \\ &= \frac{2 \times \left(-\log_2 \left(\frac{2.76 \times 10^6}{10^{12}}\right)\right)}{-\log_2 \frac{7.18 \times 10^6}{10^{12}} - \log_2 \frac{27 \times 10^6}{10^{12}}} = 1.14 \end{aligned}$$

As the similarity between *chihuahua* and *persian cat* has been incorrectly assessed (with a value above 1), we will erroneously conclude that

$$sim_{lin}\_IR\,(terrier, chihuahua) < sim_{lin}\_IR\,(chihuhaua, persian\_cat)$$

Applying IC_IR to Jiang & Conrath dissimilarity measure, the same problem appears:

$$\begin{aligned} dis_{jcn}\_IR\,(terrier, chihuahua) &= (IC\_IR\,(terrier) + IC\_IR\,(chihuahua)) \\ &\quad -2 \times IC\_IR\,(dog) \\ &= 9.45 \\ dis_{jcn}\_IR\,(chihuahua,\ persian\_cat) &= (IC\_IR\,(chihuahua) + IC\_IR\,(persian\_cat)) \\ &\quad -2 \times IC\_IR\,(mammal) \\ &= -4.66 \end{aligned}$$

In this case, the incorrectly assessed dissimilarity between *Chihuahua* and *Persian cat* results in a negative value, erroneously concluding that

$$dis_{jcn}\_IR\,(terrier, chihuahua) > dis_{jcn}\_IR\,(chihuahua,\ persian\_cat)$$

3 Contextualized information content: computing IC from the Web in a scalable and coherent manner

In order to avoid the problems presented in the previous section we can try to redefine the way in which concept probabilities for IC computation are estimated from the Web. Applying Resnik's approach to the Web (i.e. recursively adding specialized concept appearances to their subsumers) as shown in Section 2.1, introduces problems about scalability (i.e. a large number of web queries is required), as well as, a heavy dependence to both the ontology and corpus modifications. In order to tackle those issues, in this section we present a new way to coherently compute concept's IC from word's web hit counts for similarity assessment using a reduced number of queries.

We propose to compute concept appearance probabilities from web hit counts in a Web-scalable manner by contextualizing concept's term appearances in the scope of its subsumer. The hypothesis is that the hit count, at a Web scale, of the explicit co-occurrence of a word and an appropriate subsumer provides better concept appearance probabilities which enable a more accurate similarity assessment.

From the technical point of view, web search engines natively support word co-occurrences from especially formulated queries (using logic operators such as AND or +). Using this feature, we force the co-occurrence between the subsumer (e.g. *mammal*) and each of the subsumed terms (e.g. *dog*) in the web query ensuring that the IC_IR of the subsumed term (computed as *hits(dog AND mammal)*) is higher than its subsumer (computed as *hits(mammal)*). It is important to note that, in the case in which a concept is represented by several words (e.g. *persian cat*), double quotes should be used to maintain the context.

In addition to ensure the taxonomical coherence, this approach also aids to minimize ambiguity of absolute word appearances by contextualizing the search. For example, computing the occurrence of the term *dog* (referred as an *mammal*) in a corpus may give an idea of the word's appearance probability considering all its possible senses (i.e. associated concepts like *animal*, but also *fast food*); however, forcing the occurrence of *dog* and *mammal* (being *mammal* the LCS of *dog* and another concept such as *cat*) will introduce additional contextual information about the preferred word sense. Obviously, this implies a reduction of the corpus evaluated for the statistical assessment (i.e. only explicit co-occurrences are considered) and a subestimation of the real concept probability. Certainly, there will be many documents referring to the concept of *dog as a mammal* which will not explicitly include the word *mammal* in the text. However we hypothesize that, on one hand, considering the enormous size of the Web, data sparseness problems are minimized (Brill 2003). On the other hand, from the similarity computation point of view, the comparison of subestimated probabilities of the concepts will lead to more accurate assessments than probabilities based on absolute word occurrences.

Using this approach, we consider that each document of a corpus is typically using each word (which represents a web *hit* in a search engine) unambiguously. Disambiguation of term appearances at a document level is based on the observation that words tend to exhibit only one sense in a given discourse or document (context). This fact was tested by Yarowsky (1995) on a large corpus (37.232 examples), obtaining a very high precision (around 99%).

From the similarity computation point of view, we propose that the subsumer used to contextualize web queries is the LCS of the pair of evaluated concepts in a given taxonomy. In this manner, we define the *Web-based Contextualized Information Content (CIC_IR)* for a pair of concepts as follows:

Definition 2 For any pair of concepts a and b contained in a taxonomy T, the *Web-based Contextualized Information Content (CIC$_T$_IR)* of a with respect to b is:

$$CIC_T\_IR\,(a_b) = -\log_2 p_{web}\,(a_b) = -\log_2 \frac{hits\,(a\ AND\ LCS_T\,(a,b))}{total\_webs}, \tag{7}$$

where and $p_{web}\,(a_b)$ is the subestimated probability of concept a in the Web when computing its similarity against b. This probability is computed from the web hit counts returned by a search engine—*hits*—when querying the terms a and $LCS_T(a,b)$ (extracted from the taxonomy T which contains a and b) at the same time (using *AND* or '+' logic operators). *Total_webs* is the total number of resources indexed by the web search engine.

Equally, for b with respect to a:

$$CIC_T\_IR\,(b_a) = -\log_2 p_{web}\,(b_a) = -\log_2 \frac{hits\,(b\ AND\ LCS_T\,(a,b))}{total\_webs} \tag{8}$$

As stated above, this is a subestimation of concept's probability. Note that the presented formula is different to the conditioned probability of the term with respect to the *LCS* (i.e. $p(a|LCS(a,b)) = hits(a\ AND\ LCS(a,b))/hits(LCS(a,b))$). The conditioned probability calculation, due to the denominator, will introduce the recursive problem of LCS concept probability estimation from absolute word hit counts, which we try to avoid.

As stated above, with the proposed approach, we ensure that:

Proposition 1 *The IC_IR of the subsumer is always inferior to the CIC$_T$_IR of its subsumed terms.*

$$IC\_IR\,(LCS\,(a,b)) \le \min\,(CIC_T\_IR\,(a_b)\,, CIC_T\_IR\,(b_a)) \tag{9}$$

This guarantees that the subsumer will be more general -less informative- than its specializations, because the latter's ICs are computed in the context of the documents covering the subsumer. In consequence, from the similarity computation point of view, IC values will be taxonomically coherent.

It is important to note that, with this method, only one web query is needed to estimate the IC of each evaluated concept. So, the cost for a given pair of concepts with one LCS in common is constant. In addition, modifications in the taxonomy, which may affect Resnik-like IC computation (like adding a new sibling to the taxonomic specialization of a given subsumer), does not influence the calculation of CIC_T_IR. In consequence, our approach is more scalable and more independent to changes in the knowledge base.

3.1 Introducing CIC_T_IR to Resnik-based measures

As Resnik similarity measure only considers the occurrence of the LCS in a corpus and not the IC of the evaluated concepts, CIC_T_IR cannot be directly applied. For measures like Lin or Jiang & Conrath, which evaluate the difference between the IC of subsumed terms against their LCS (see Section 2), the introduction of CIC_T_IR can aid to obtain a taxonomically coherent, less ambiguous an more accurate similarity assessment from the Web. More details will be given in the evaluation section. The proposed contextualized versions of Lin and Jiang & Conrath functions are defined below.

Definition 3 The Web-based contextualized version of the Lin similarity measure ($sim_{lin}\_CIC_T\_IR$) between concepts a and b contained in the taxonomy T is defined as follows:

$$\begin{aligned} & sim_{lin}\_CIC_T\_IR\,(a,b) \\ & = \frac{2\times IC\_IR\,(LCS_T\,(a,b))}{(CIC_T\_IR\,(a_b)+CIC_T\_IR\,(b_a))} \\ & = \frac{2\times\left(-\log_2\dfrac{hits\,(LCS_T\,(a,b))}{total\_webs}\right)}{\left(-\log_2\dfrac{hits\,(a\ AND\ LCS_T\,(a,b))}{total\_webs}-\log_2\dfrac{hits\,(b\ AND\ LCS_T\,(a,b))}{total\_webs}\right)} \end{aligned} \quad (10)$$

Proposition 2 *The modified function, $sim_{lin}\_CIC_T\_IR$ is a similarity measure because it fulfills the following properties (Euzenat and Shvaiko 2007):*

$$\forall a,b\in \mathrm{O},\, sim\,(a,b)=sim\,(b,a) \qquad (symmetry) \qquad (11.1)$$

$$\forall a,b,c\in \mathrm{O},\, sim\,(a,a)\geq sim\,(b,c) \qquad (maximality) \qquad (11.2)$$

$$\forall a,b\in \mathrm{O},\, sim\,(a,b)\geq 0 \qquad (positiveness) \qquad (11.3)$$

Proof Due to the original Lin measure accomplishes all these properties (Lin 1998), by definition (as $\mathrm{LCS}(a, b) = \mathrm{LCS}(b, a)$ and all ratios will return values between 0 and 1), $sim_{lin}\_CIC_T\_IR$ also accomplishes the properties of *symmetry* (11.1) and *positiveness* (11.3). *Maximality* (11.2) is also accomplished considering that the LCS for the pair of concepts a and a is the same a. So, $\mathrm{CIC_T\_IR}(a_a) = \mathrm{IC\_IR}(a)$ that is the original Lin function, which fulfills the *maximality* property. □

Definition 4 The Web-based contextualized version of the Jiang & Conrath measure ($dis_{lin}\_CIC_T\_IR$) for concepts a and b contained in the taxonomy T is defined as follows:

$$\begin{aligned}
&dis_{jcn}\_CIC_T\_IR\,(a, b)\\
&\quad= (CIC_T\_IR\,(a_b) + CIC_T\_IR\,(b_a)) - 2 \times IC\_IR\,(LCS_T\,((a, b))\\
&\quad= \left(-\log_2 \frac{hits\,(a\ AND\ LCS_T\,((a, b))}{total\_webs} - \log_2 \frac{hits\,(b\ AND\ LCS_T\,((a, b))}{total\_webs}\right)\\
&\qquad - 2 \times \left(-\log_2 \frac{hits\,(LCS_T\,((a, b))}{total\_webs}\right)
\end{aligned} \tag{12}$$

Proposition 3 *The function* $dis_{jcn}\_\mathrm{CIC_T\_IR}$ *fulfills the properties of dissimilarity measures (Euzenat and Shvaiko 2007):*

$$\forall a, b \in \mathrm{O}, dis\,(a, b) \geq 0 \qquad (positiveness) \tag{13.1}$$

$$\forall a \in \mathrm{O}, dis\,(a, a) = 0 \qquad (minimality) \tag{13.2}$$

$$\forall a, b \in \mathrm{O}, dis\,(a, b) = dis\,(b, a) \qquad (symmetry) \tag{13.3}$$

Proof Because of the original Jiang & Conrath measure accomplishes these properties (Jiang and Conrath 1997), by definition (as $\mathrm{LCS}(a, b) = \mathrm{LCS}(b, a)$), $dis_{jcn}\_CIC_T\_IR$ accomplishes the *symmetry* property (13.3). Being $\mathrm{CIC_T\_IR}$ a taxonomically coherent estimation (i.e. for subsumed terms, it will be always higher than the IC_IR of the LCS), then $(\mathrm{CIC_T\_IR(a_b)} + \mathrm{CIC_T\_IR(b_a)}) > 2{*}\mathrm{IC\_IR(a,b)}$, accomplishing the *positiveness* property (13.1). *Minimality* (13.2) is also fulfilled considering that the LCS for the pair of concepts a and a is the same a. So, $\mathrm{CIC_T\_IR}(a_a) = \mathrm{IC\_IR}(a)$ results in the original dis_{jcn} function, which also accomplishes the *minimality* property. □

3.2 Dealing with polysemy and synonymy

Typical domain ontologies are unambiguous (Dujmovic and Bai 2006) (i.e. a unique LCS represented by one textual form is available for any pair of concepts). However, general purpose ontologies, such as WordNet, typically implement *polysemy* by representing several *is-a* relationships for the same concept and *synonymy* by associating a list of semantically equivalent terms to each sense (*synsets*). In the former case, several LCS may exist for different taxonomical classifications of a given pair of terms; in the latter case, several textual forms for each LCS may be available.

Resnik-like measures tackle polysemy by using the *Most Specific Common Subsumer* (MSCS) which corresponds to the LCS with the highest IC value (Resnik 1995) (i.e. for a pair of terms, they consider the pair of most similar senses represented by the MSCS). They take all the possible subsumers, compute the similarity for each of them and take the maximum value as final result (the minimum for dissimilarity measures). Synonyms associated to LCSs are not a problem in their approach because the background corpus used by those measures incorporates frequencies of concepts (WordNet *synsets*) rather than words.

In the framework proposed in this paper, in which a general ontology and corpus can be also used, these two issues must be considered. For polysemic cases the strategy will be the same as Resnik: all the LCSs available through the several taxonomic paths are retrieved, the similarity measure is computed for each of them and highest value (or lowest for dissimilarity) is taken.

In the case of synonyms (i.e. different textual forms are available for the same concept) one may consider to add the hit counts for the queries constructed with the available LCS synonyms. For example, being *dog* and *canis* synonyms of the subsumer of *terrier*, we can compute *hits(terrier AND dog NOT canis)* + *hits(terrier AND canis NOT dog)* + *hits(terrier AND canis AND dog)*. However, in cases with a large set of synonyms (which is common in WordNet), a large amount of queries are needed, because they must include all the possible synonym combinations, as well as, a considerable number of keywords (resulting in a query which length may be not supported by typical web search engines). In addition, the final value will accumulate a considerable error derived from the individual errors inherent to the *estimated* hit counts provided by the search engine. Finally, this will make the similarity results dependant on the synonym coverage of each concept. Instead, we opted to consider each LCS synset synonym individually, computing the similarity value for each one and taking as a result the highest one (the lowest for dissimilarity measures). In this way, the LCS would correspond to the word that best contextualizes the queries (i.e. the less ambiguous textual form). During the research, we observed that this strategy leads to more accurate results than considering the sum of synonyms hit counts.

Definition 5 The generalized version of the sim$_{\mathrm{lin}}$_CIC$_{\mathrm{T}}$_IR for the case of multiple subsumers and textual forms available in the taxonomy T is defined as follows:

$$sim_{lin}\_CIC_T\_IR\,(a,b) = \max_{L\in S(a,b)}\left(\frac{2\times\left(-\log_2\frac{hits\,(L)}{total\_webs}\right)}{\left(-\log_2\frac{hits\,(a\ AND\ L)}{total\_webs}-\log_2\frac{hits\,(b\ AND\ L)}{total\_webs}\right)}\right), \tag{14}$$

where $S(a,b)$ is the set of textual forms (synonyms) of all the LCS that subsume a and b in the given taxonomy T.

Definition 6 The generalized version of the dis_{jcn}_CIC_T_IR measure for the case of multiple subsumers and textual forms available in the taxonomy T is defined as follows:

$$dis_{lin}\_CIC_T\_IR\,(a,b) = \min_{L\in S(a,b)}\left(\left(-\log_2\frac{hits\,(a\ AND\ L)}{total\_webs} - \log_2\frac{hits\,(b\ AND\ L)}{total\_webs}\right) - 2\times\left(-\log_2\frac{hits\,(L)}{total\_webs}\right)\right), \qquad (15)$$

where $S(a,b)$ is the set of textual form (synonyms) of all the LCS that subsume a and b in the given taxonomy T.

3.3 An example

In this section, the proposed contextualized versions of Lin and Jiang & Conrath similarity functions are tested with the example introduced in Section 2.2.

In that example, the similarity between *Terrier* and *Chihuahua* was compared to the one between *Chihuahua* and *Persian Cat*. Note that the former should be more similar than the second, because they both are dogs. The information provided by the taxonomy presented in Fig. 1 has been used to contextualize the queries as proposed by CIC_T_IR. The results obtained are the following:

$$sim_{lin}\_CIC_T\_IR\,(terrier, chihuahua) = \frac{2\times\left(-\log_2\frac{hits\,(dog)}{total\_webs}\right)}{\left(-\log_2\frac{hits\,(terrier\ AND\ dog)}{total\_webs} - \log_2\frac{hits\,(chihuahua\ AND\ dog)}{total\_webs}\right)} = \frac{2\times\left(-\log_2\left(\frac{204\times 10^6}{10^{12}}\right)\right)}{-\log_2\frac{3.86\times 10^6}{10^{12}} - \log_2\frac{2.79\times 10^6}{10^{12}}} = 0.67$$

$$sim_{lin}\_CIC_T\_IR\,(chihuahua, persian\_cat) = \frac{2\times\left(-\log_2\frac{hits\,(mammal)}{total\_webs}\right)}{\left(-\log_2\frac{hits\,(chihuahua\ AND\ mammal)}{total\_webs} - \log_2\frac{hits\,(\text{"}persian\_cat\text{"}\ AND\ mammal)}{total\_webs}\right)} = \frac{2\times\left(-\log_2\left(\frac{2.76\times 10^6}{10^{12}}\right)\right)}{-\log_2\frac{55400}{10^{12}} - \log_2\frac{6290}{10^{12}}} = 0.719$$

Although in this case the occurrence probabilities are taxonomically coherent and similarity differences have been greatly reduced, an erroneous conclusion is again reached:

$$sim_{lin}\_CIC_T\_IR\,(terrier, chihuahua) < sim_{lin}\_CIC_T\_IR\,(chihuhaua, persian\_cat)$$

For Jiang & Conrath modified measure, the dissimilarity is now properly assessed.

$$\begin{aligned} &dis_{jcn}\_CIC_T\_IR\,(terrier, chihuahua) \\ &\quad = \left(-\log_2 \frac{hits\,(terrier\ AND\ dog)}{total\_webs} - \log_2 \frac{hits\,(chihuahua\ AND\ dog)}{total\_webs}\right) \\ &\qquad - 2 \times \left(-\log_2 \frac{hits\,(dog)}{total\_webs}\right) = 11.93 \end{aligned}$$

$$\begin{aligned} &dis_{jcn}\_CIC_T\_IR\,(chihuahua, persian\_cat) \\ &= \left(-\log_2 \frac{hits\,(chihuahua\ AND\ mammal)}{total\_webs} - \log_2 \frac{hits\,(\text{"}persian\_cat\text{"}\ AND\ mammal)}{total\_webs}\right) \\ &\quad - 2 \times \left(-\log_2 \frac{hits\,(mammal)}{total\_webs}\right) = 14.42 \end{aligned}$$

$$dis_{jcn}\_CIC_T\_IR\,(terrier, chihuahua) > dis_{jcn}\_CIC_T\_IR\,(cat, dog)$$

In these tests, the performance of both measures has been greatly improved by the inclusion of $\mathrm{CIC_T\_IR}$ because, even though concept probabilities have been subestimated, they are based in less ambiguous Web occurrences. However, the problem of estimating the IC_IR of the LCS in an uncontextualized manner remains. Concept probabilities will be taxonomically coherent but problems presented in Section 2.1 may affect the LCS probability estimation. In consequence, the similarity may be not properly measured, like in the first case.

In order to circumvent this problem, in the next section, collocation measures are considered. As presented in the introduction, those measures only consider concept occurrence and not their LCS alone. In consequence, they will be immune to that issue providing better performance, as will be shown in the evaluation section.

4 Analyzing collocation measures

As stated in the introduction, there exist other measures which seek for the co-occurrence between terms in order to estimate their correlation. In this case, they are completely unsupervised, as no background knowledge (a part from an unprocessed corpus) is employed. Those measures have been applied in similarity estimation based on the relation that exists between term co-occurrence in a corpus and their similarity (Spence and Owens 1990).

In order to statistically assess the degree of correlation and, as stated above, the similarity between words, standard collocation functions have been proposed. Formally, they are defined in the following way:

$$c_k(a,b) = \frac{p(ab)^k}{p(a)\,p(b)}, \tag{16}$$

being $p(a)$ the probability that the word a occurring within the text and $p(ab)$ the probability of co-occurrence of words a and b. Here, the collocation of a and b is defined as the comparison between the probability of observing a and b together with respect to observing them independently. If a and b are statistically independent, the probability that they co-occur is given by the product $p(a)p(b)$. If they are not independent, and they have a tendency to co-occur in a corpus, $p(ab)$ will be greater than $p(a)p(b)$. Therefore the ratio between $p(ab)$ and $p(a)p(b)$ is a measure of the degree of statistical dependence between a and b (Turney 2001).

The most typical forms of collocation functions are the *Symmetric Conditional Probability* (SCP), defined as c_2 (Ferreira da Silva and Lopes 1999) and the *Pointwise Mutual Information* (PMI), defined as $log_2 c_1$ (Church et al. 1991). In the latter case, the measure can be expressed in terms of the IC for a when we observe b and IC for b when we observe a (17).

$$PMI(a,b) = \log_2 \frac{p(ab)}{p(a)\,p(b)} = (IC(a) + IC(b)) - IC(ab) \tag{17}$$

Considering the Web as a valuable corpus from which compute reliable statistics about information distribution, PMI was adapted by Turney (2001) to approximate concept probabilities using the hit counts of a web search engine. The equation is specified as follows:

$$PMI\_IR(a,b) = \log_2 \frac{\dfrac{hits(a\ \ AND\ \ b)}{total\_webs}}{\dfrac{hits(a)}{total\_webs} \times \dfrac{hits(b)}{total\_webs}} \tag{18}$$

However, from previous investigations (Downey et al. 2007), SCP have outperformed PMI by a large margin in the task of assessing similarity values for pairs of words, as it is less dependent on the order of magnitude of occurrence values. In the same manner as Turney, SCP can be adapted (19) to compute concept probabilities from web hit counts (Downey et al. 2007).

$$SCP\_IR(a,b) = \frac{\left(\dfrac{hits(a\ \ AND\ \ b)}{total\_webs}\right)^2}{\dfrac{hits(a)}{total\_webs} \times \dfrac{hits(b)}{total\_webs}} = \frac{(hits(a\ \ AND\ \ b))^2}{hits(a) \times hits(b)} \tag{19}$$

Even though both measures have been applied to the task of evaluating concept relatedness (Downey et al. 2007; Etzioni et al. 2005), due to their lack of semantics, they offer a limited performance (Lemaire and Denhière 2006). This is caused by the inaccurate concept probability estimation from absolute word hit counts. In addition, decontextualized term co-occurrences in a document may be indicative of *relatedness* (Patwardhan and Pedersen 2006) but not necessarily of *semantic similarity* (Lemaire and Denhière 2006).

4.1 An example

In order to illustrate the presented issues, let us consider the Web-based SCP_IR value of taxonomic siblings (with respect to *mammals*) such as *cat* and *dog*. Following the example in Fig. 1, we have,

$$SCP\_IR\,(cat, dog) = \frac{(hits\,(cat\ \ AND\ \ dog))^2}{hits\,(cat) \times hits\,(dog)} = \frac{\left(41.3 \times 10^6\right)^2}{511 \times 10^6 \times 204 \times 10^6} = 0.016$$

However, computing the SCP_IR of semantically farther concepts such as *dog* and *food*, we obtain the following score:

$$SCP\_IR\,(dog,\ food) = \frac{(hits\,(dog\ \ AND\ \ food))^2}{hits\,(dog) \times hits\,(food)} = \frac{\left(69.5 \times 10^6\right)^2}{204 \times 10^6 \times 477 \times 10^6} = 0.049$$

As a result, we will erroneously conclude that

$$SCP\_IR\,(cat, dog) < SCP\_IR\,(dog,\ food)$$

This is a case in which word co-occurrences in a corpus (as studied in Section 2.1) are not directly proportional to concepts' similarity. In this case, *dog* and *food* relationship is biased by *dog food companies* and also by the fact that *dog* in the sense of *food—hot dog*—is commonly used. As stated above, some kind of relationship between concepts, like *hyponymy*, *meronymy*, *antonymy* or any other kind of *non-taxonomic* relationship, cannot be assessed by absolute co-occurrence due to the lack of semantic content (Lemaire and Denhière 2006).

5 Applying CIC$_T$_IR to collocation measures

In order to overcome the presented problems of Web-based collocation measures due to their lack of semantics, in this section we will modify them by estimating concept probabilities by means of the taxonomy-based CIC$_T$_IR presented in Section 3. The goal is to improve the performance of those measures by considering the additional knowledge provided by a taxonomic structure, exploiting the LCS to contextualize queries. Even though modified versions cannot be considered as unsupervised due to the need of a background taxonomy, they offer an alternative to Resnik-like semantic measures, avoiding some of their problems (introduced in Section 3.3).

In this case, on the contrary to Resnik-like measures, the explicit co-occurrence of evaluated concepts is involved. In order to properly assess this concept co-occurrence from the Web in a contextualized manner, we extend the CIC$_T$_IR definition (Definition 2) in the following way:

Definition 7 For any pair of concepts a and b contained in a taxonomy T, the *Web-based Contextualized Information Content (CIC$_T$_IR)* of the co-occurrence of *a and b* is:

$$\begin{aligned} CIC_T\_IR\,(ab) &= -\log_2 p_{web}\,(ab) \\ &= -\log_2 \frac{hits\,(a\ AND\ b\ \ AND\ \ LCS_T\,(a, b))}{total\_webs}, \end{aligned} \tag{20}$$

where $p_{web}(ab)$ is the probability of co-occurrence of concepts a, b, estimated from the co-occurrence of words a, b and $LCS_T(a, b)$ in the Web, which is computed from the web hit counts of a web search engine.

In addition to the advantages introduced in Section 3, this function permits to minimize term co-occurrence ambiguity, using the information of the taxonomic structure. In fact, co-occurrence will be biased by the ontological knowledge to the taxonomical side due to the additional semantics provided by the inclusion of the subsumer. In consequence, non-taxonomic relationships between the evaluated terms (which may hamper the similarity estimation as stated in the previous section) will have less weight in the statistical assessment.

In order to introduce CIC$_\text{T}$_IR to collocation measures, we have rewritten the classical collocation definition (16) in terms of IC by including the log_2 function (Definition 8). Considering that term occurrence follows a hyperbolic distribution (Hotho et al. 2002) and high order occurrences tend to overestimate similarity (Lemaire and Denhière 2006), the logarithm function (being a monotonic function) helps to smooth absolute occurrence values without altering the value tendency.

Definition 8 Given the concepts a and b, the collocation measure expressed in terms of concept's IC is defined as follows:

$$c_k\_IC(a, b) = \log_2 \frac{p(ab)^k}{p(a)\,p(b)} = (IC(a) + IC(b)) - k \times IC(ab) \tag{21}$$

This new version can be directly contextualized by means of the CIC$_\text{T}$_IR calculation, which takes into account the information provided by the taxonomy T that includes a and b.

Definition 9 Given the concepts a and b, their *Web-based Contextualized Collocation measure* is defined as:

$$\begin{aligned} & c_k\_CIC_T\_IR(a, b) \\ & = (CIC_T\_IR(a_b) + CIC_T\_IR(b_a)) - k \times CIC_T\_IR(ab) \\ & = \log_2 \frac{\left(\dfrac{hits(a \;\; AND \;\; b \;\; AND \;\; LCS_T(a, b))}{total\_webs}\right)^k}{\dfrac{hits(a \;\; AND \;\; LCS_T(a, b))}{total\_webs} \times \dfrac{hits(b \;\; AND \;\; LCS_T(a, b))}{total\_webs}} \end{aligned} \tag{22}$$

As a result, similarity assessment from concepts' probabilities will be only based on observations with minimized ambiguity resulting in more reliable—even subestimated—occurrence values. It is important to note that, in this definition, all the Web queries have been contextualized (a goal which, for example, could not be achieved using concept conditional probabilities with respect to the LCS, as stated in Section 3). Consequently, this measure will not have the problems found in the Resnik-like measures, related to ambiguous estimation of the IC of the LCS (discussed in Section 3.3).

As explained in Section 4, *log* c_1 (PMI) and c_2 (SCP) are two common forms of c_k that are used for concept similarity estimation. Using the presented notation, $c_1\_CIC_T\_IR$ will correspond to $PMI\_CIC_T\_IR$ (Definition 10) and $c_2\_CIC_T\_IR$ will correspond to *SCP_CIC$_T$_IR* (Definition 11). It is expected that the SCP will have better results also in this version as it has offered the best performance in its original form (Downey et al. 2007).

Definition 10 PMI computation when introducing CIC$_{\mathrm{T}}$_IR for concepts a and b in the taxonomy T is defined as follows:

$$PMI\_CIC_T\_IR\,(a,b) = \log_2 \frac{\dfrac{hits\,(a\ \ AND\ \ b\ \ AND\ \ LCS_T\,(a,b))}{total\_webs}}{\dfrac{hits\,(a\ \ AND\ \ LCS_T\,(a,b))}{total\_webs} \times \dfrac{hits\,(b\ \ AND\ \ LCS_T\,(a,b))}{total\_webs}} \tag{23}$$

Definition 11 IC-based SCP computation when introducing CIC$_{\mathrm{T}}$_IR for concepts a and b in the taxonomy T is defined as follows:

$$SCP\_CIC_T\_IR\,(a,b) = \log_2 \frac{(hits\,(a\ \ AND\ \ b\ \ AND\ \ LCS_T\,(a,b)))^2}{hits\,(a\ \ AND\ \ LCS_T\,(a,b)) \times hits\,(b\ \ AND\ \ LCS_T\,(a,b))} \tag{24}$$

Note that the *total_webs* constant is simplified as it is common to the numerator and denominator.

It is important to mention that the presented measures are different to the general collocation function when considering correlations between three terms (which, in our case may correspond to a, b and *LCS(a,b)*) using the generalized version of the function (25) (Downey et al. 2007). In that case, it does not takes into account the additional knowledge introduced by the fact that LCS subsumes both terms at the same time.

$$g_k\,(a,b,c) = \frac{p\,(abc)^k}{p\,(a)\,p\,(b)\,p\,(c)} \tag{25}$$

5.1 Properties

In this section, the properties that should fulfill a similarity measure (as introduced in Section 3.1) are studied for the proposed collocation functions.

Proposition 4 *The function, c_k_*CIC$_{\mathrm{T}}$_IR *is a similarity measure accomplishing symmetry* (11.1), *maximality* (11.2) *and positiveness* (11.3) *properties.*

Proof The *symmetry* property is fulfilled because all operations are commutative, *LCS(a,b)* = *LCS(b,a)* and unquoted web queries like *hits(a AND b)* and *hits(b AND a)* give the same results when performed over a web search engine (more details in the evaluation section).

Considering the presence of the log_2 function, *positiveness* is accomplished when the relation between the numerator and denominator is equal or higher than 1. For

PMI_CIC$_T$_IR this is always true because, as terms in a corpus are not statistically independent (as introduced in Section 4), the numerator will be always equal or higher than the denominator. For *SCP_CIC$_T$_IR* the situation is the contrary, returning, for the maximal case (*sim(a,a)*) a value of *log*$_2$*((hits(a))*^{2}*/(hits(a)*hits(a))* = *log*$_2$1. So, for the most similar case, the measure will return a 0 value. Less similar pairs will result in lower numerator values with respect to the denominator and, in consequence, in results in the range of *(log*$_2$0, *log*$_2$1*]*. As a result, the final similarity value will belong to the interval *(*− ∞,0*]*. As similarity values are negative, *positiveness property* (11.3) is not accomplished for *SCP_CIC$_T$_IR*.

The *maximality* property states that the most similar concept (the highest value) to *a* must be *a*. Considering that the LCS for the pair of concepts *a* and *a* is also *a* and that, in a web search engine, the number of times that the same word appears in a unquoted web query does not affect the results (i.e. *hits(a AND a)* = *hits(a)*), we have that equation can be rewritten as:

$$c_k\_CIC_T\_IR\,(a,a) = \log_2 \frac{\left(\dfrac{hits\,(a)}{total\_webs}\right)^k}{\dfrac{hits\,(a)}{total\_webs} \times \dfrac{hits\,(a)}{total\_webs}}$$

As in that case, the numerator value is maximum for concept *a* (i.e. *hits(a)* will be higher than any other query involving *a*), *maximality property* (11.2) is also accomplished. This function gives the maximum value of 1 for *PMI_CIC$_T$_IR*, and the maximum value of 0 for the *SCP_CIC$_T$_IR*. □

Regarding the negative values of the *SCP_CIC$_T$_IR* measure, the sign of the equation can be changed, transforming the similarity measure into a dissimilarity (Definition 12) with positive values.

Definition 12 Being *a* and *b* two concepts contained in the taxonomy *T*, SCP-based dissimilarity function when introducing CIC$_T$_IR is defined as follows:

$$\begin{aligned} &Dis_{SCP\_CIC_T\_IR}\,(a,b) \\ &\quad = -\log_2 \frac{(hits\,(a\ \ AND\ \ b\ \ AND\ \ LCS_T\,(a,b)))^2}{hits\,(a\ \ AND\ \ LCS_T\,(a,b)) \times hits\,(b\ \ AND\ \ LCS_T\,(a,b))} \end{aligned} \qquad (26)$$

Proposition 5 *The function, Dis$_{SCP\_CICT\_IR}$ is a dissimilarity measure accomplishing positiveness* (13.1)*, minimality* (11.2) *and symmetry* (11.3) *properties.*

Proof As a dissimilarity measure, *positiveness* (13.1) is accomplished as, when changing the sign, the value range of the results will belong to the interval *[*0,+∞*)*. Likewise, *minimality* (13.2) is also true as, for identical pairs the function will return a 0 value. *Symmetry* (13.3) is accomplished by definition. □

5.2 Ontologies with polysemy and synonymy

When considering ontologies modeling polysemy and synonymy, the same strategy presented in Section 3.2 is proposed for the case of collocation measures. So, generalized versions of the collocation measures are defined (Definition 13). They consider all the common subsumers and textual forms, compute all similarity (or dissimilarity) values and take the highest (or lowest) one.

Definition 13 The generalized version of the collocation-based functions for the case of multiple subsumers and synonyms available in the taxonomy T is defined as follows:

$$c_k\_CIC_T\_IR\,(a,b) = \max_{L \in S(a,b)} \left(\log_2 \frac{\left(\dfrac{hits\,(a \;\; AND \;\; b \;\; AND \;\; L)}{total\_webs}\right)^k}{\dfrac{hits\,(a \;\; AND \;\; L)}{total\_webs} \times \dfrac{hits\,(b \;\; AND \;\; L)}{total\_webs}} \right), \quad (27)$$

where $S(a,b)$ is the set of textual form (synonyms) of all the LCS that subsume a and b in the given taxonomy T.

Note that for the case of $Dis_{SCP\_CICT\_IR}$, being a dissimilarity measure, the minimum value should be considered.

5.3 An example

The same example studied in Section 2.2 is now applied to the proposed collocation-based measures. In particular, the results of the dissimilarity $Dis_{SCP\_CICT\_IR}$ are presented. Being *terrier* and *chihuahua* specializations of *dog* and *chihuahua* and *persian cat* specializations of *mammal*, we now obtain the following values:

$$\begin{aligned} & Dis_{SCP\_CIC_T\_IR}\,(terrier, chihuahua) \\ & = -\log_2 \frac{(hits\,(terrier \;\; AND \; chihuahua \;\; AND \;\; dog))^2}{hits\,(terrier \;\; AND \;\; dog) \times hits\,(chihuahua \;\; AND \;\; dog)} \\ & = -\log_2 \frac{935{,}000^2}{3{,}860{,}000 \times 2{,}790{,}000} = 3.625 \end{aligned}$$

$$\begin{aligned} & Dis_{SCP\_CIC_T\_IR}\,(chihuahua, persian\_cat) \\ & = -\log_2 \frac{(hits\,(chihuahua \; AND \; "persian \; cat" \;\; AND \;\; mammal))^2}{hits\,(chihuahua \; AND \; mammal) \times hits\,("persian \; cat" \; AND \; mammal)} \\ & = -\log_2 \frac{531^2}{55{,}400 \times 6{,}290} = 10.27 \end{aligned}$$

In this case, as ambiguity is minimized because all the queries have been contextualized in the scope of the LCS, we will correctly conclude that

$$Dis_{SCP\_CIC_T\_IR}\,(chihuahua, terrier) < Dis_{SCP\_CIC_T\_IR}\,(chihuahua, persian\_cat)$$

Applying the same measure to the case presented in Section 4.1 (being *cat* and *dog* specializations of *mammal* and *cat* and *food* specializations of *entity*), we obtain the following values:

$$\begin{aligned} & Dis_{SCP\_CIC_T\_IR}(cat, dog) \\ & = -\log_2 \frac{(hits\,(cat \;\; AND \; dog \;\; AND \;\; mammal))^2}{hits\,(cat \;\; AND \;\; mammal) \times hits\,(dog \;\; AND \;\; mammal)} \\ & = -\log_2 \frac{569{,}000^2}{906{,}000 \times 933{,}000} = 1.39 \end{aligned}$$

$$\begin{aligned} & Dis_{SCP\_CIC_T\_IR}(dog,\ food) \\ & = -\log_2 \frac{(hits\,(dog \;\; AND \;\; food \;\; AND \;\; entity))^2}{hits\,(dog\ AND\ entity) \times hits\,(food\ AND\ entity)} \\ & = -\log_2 \frac{986{,}000^2}{2{,}040{,}000 \times 6{,}490{,}000} = 3.775 \end{aligned}$$

Again, as the LCS helps to focus the statistical assessment of co-occurrence towards the taxonomic side and minimize the ambiguity of the word *dog* (in the *mammal* sense), it is correctly concluded that

$$Dis_{SCP\_CIC_T\_IR}(cat, dog) \; < \; Dis_{SCP\_CIC_T\_IR}(dog,\ food)$$

6 Evaluation

An effective way of evaluating computerized similarity assessments consists on comparing them to human judgments for the same set of terms. Computing the *correlation* between the computerized and human-based ratings, we are able to obtain a quantitative value of the quality of the similarity function. This enables an objective comparison against other measures.

As a test bed, Miller and Charles (1991) proposed a set of words for which human subjects have rated their similarity from 0 to 4. Many authors (Resnik 1995; Jiang and Conrath 1997; Lin 1998) have used this benchmark to compare the performance of their measures. Concretely, Resnik (1995) replicated the example providing more accurate ratings than Miller and Charles, because a group of experts (instead of regular subjects) were requested to rate the similarity of the pairs of words. The correlation of the Resnik's experiment was 0.884, which represents an upper bound to what one could expect from a machine-based similarity assessment. In this experiment, only word pairs contained in WordNet where considered (28 from the original set of 30).

In our test, we have taken the same set of 28 word pairs, as shown in Table 2, and their averaged expert ratings provided by Resnik. As background ontology from which extract LCS for word pairs, we use the latest version of WordNet (3.0).[6]

[6] http://wordnet.princeton.edu/

Table 1 Hit count returned by Google and Bing for equivalent queries [accessed: May 26th, 2009]

Terms to evaluate	Equivalent web queries	Google Web hit count	Bing Web hit count
dog cat	dog cat	173,000,000	72,600,000
	cat dog	28,800,000	70,300,000
	"dog" "cat"	30,900,000	72,600,000
	dog AND cat	26,100,000	72,600,000
dog dog	dog dog	449,000,000	208,000,000
	dog AND dog	374,000,000	208,000,000
	dog	396,000,000	209,000,000
	"dog"	332,000,000	208,000,000
cat dog mammal	cat dog mammal	277,000	519,000
	dog cat mammal	403,000	519,000
	cat mammal dog	1,740,000	519,000
	dog mammal cat	1,740,000	519,000
	mammal dog cat	404,000	516,000
	mammal cat dog	277,000	519,000

Finally, in order to obtain term appearances from the Web we use Bing[7] as the search engine.

Even though other search engines can be used (Google, for example, offers a higher IR recall (Dujmovic and Bai 2006)), we found inconsistencies in the co-occurrence estimation in some of them which may produce unexpected results and compromise the similarity properties. Some examples of problematic cases for Google are provided in Table 1. In that case, quite different hit counts are obtained for equivalent web queries. Moreover, we observed a high variability in the hit counts for tests performed within a short period of time (days). Contrarily, Bing has provided consistent results during the different tests. Minimal variations in hit counts (also shown in Table 1) have been observed mainly motivated by the use of different cached data from one query to another or changes in the IR database.

In any case, other Web search engines may be also suitable if they provide coherent results. As discussed in Sánchez (2008), although the absolute occurrence values for a specific query may be quite different from one search engine to another, the final similarity values tend to be very similar as they are based in relative functions.

In our experiments, the results of the proposed modifications to Resnik-based and collocation-based similarity measures ($\mathrm{sim}_{\mathrm{lin}}$_$\mathrm{CIC}_{\mathrm{T}}$_IR, $\mathrm{dist}_{\mathrm{jcn}}$_$\mathrm{CIC}_{\mathrm{T}}$_IR, PMI_$\mathrm{CIC}_{\mathrm{T}}$_IR, $\mathrm{Dis}_{SCP\_CICT\_IR}$) have been compared against their original forms. In all cases, we have used the Web hit counts to estimate probabilities and compute concept's IC. This compares the contextualized and non-contextualized web-based concept probability assessment. The performance of each measure is evaluated by computing the correlation of the values obtained for each word pair against the human ratings employed as baseline (Resnik 1995). All the measures have been tested in the same conditions, executing the tests at the same moment (to minimize variance due to web-IR estimation changes) and, for the case of polysemic WordNet concepts, using the generalized versions presented in Sections 3.2 and 5.2.

As some of the measures involved in the test compute similarity (Resnik, Lin, PMI and SCP) and others evaluate dissimilarity (Jiang & Conrath and $\mathrm{Dis}_{\_SCP\_CICT\_IR}$),

[7] http://www.bing.com/

for consistency in comparison, we have converted all functions to similarity measures. Conversion is performed simply by changing the sign. Note that this conversion does not affect the result of the evaluation, since a linear transformation of the values will not change the magnitude of the resulting correlation coefficient, although its sign have changed from positive to negative.

Table 2 shows the complete list of results of each similarity measure for each pair of words. Results have been grouped according to the type of similarity measure. In the second column there are displayed the averaged human ratings of Resnik's study, which achieved a correlation between the experts of 0.884; these values are taken as the baseline to compare computerized approaches. The next five columns show the similarity results given by Web-based Resnik-like measures and their corresponding $\mathrm{CIC_T}$-based versions. Finally, the IC-based collocation measures including log_2 (c_1_IC_IR, which corresponds to PMI_IR, and c_2_IC_IR, which corresponds to SCP_IC_IR), can be compared against their corresponding $\mathrm{CIC_T}$-based versions. The values in italics correspond to the contextualized measures proposed in this paper. For each column, the correlation of the similarity values against the human ratings is provided as an indication of the result's quality.

Analyzing the values, in general, we can say that the results presented in the table are according to the hypothesis described during the paper. The conclusions that we can draw are the following:

- Classical Resnik-like measures perform poorly when only absolute word occurrences are used to assess concept probabilities (i.e. no tagged corpus is available). The inaccurate estimation derived from the language ambiguity and the lack of taxonomic coherence in the IC computation hamper the final results (correlation values range from 0.35 to 0.4). Lin and Jiang & Conrath are the most handicapped by the latter issue due to their explicit comparison between concept's IC and their subsumer (correlations are below 0.4). Comparatively, recovering the results obtained using the SemCor (Miller et al. 1993) as corpus (which its latest version correspond to WordNet 1.6 synsets), Resnik measure obtained a correlation of 0.794. Jiang & Conrath measure obtained a correlation among 0.794 and 0.828 according to ad-hoc predefined weighting parameters. Both are quite near to the human upper bound of 0.884 computed by Resnik replication (Resnik 1995). However, this quality is heavily associated to the accurate frequencies computed from the limited manually disambiguated corpus, tagged according to WordNet synsets (as introduced in Section 2). As shown in our tests, in lack of this tagged data gives much lower accuracy.
- The inclusion of the contextualized version of IC computation in Lin and Jiang & Conrath, due to the additional context, statistics are more accurate. As a result, they clearly outperform the basic versions, almost doubling the correlation value (0.67 vs. 0.36). In this case, even though concept probabilities have been subestimated, the monotonic coherence of IC computation with respect to the taxonomic structure and the minimized ambiguity of word occurrences certainly improve the results.
- Regarding the collocation measures, even though being unsupervised, they tend to outperform Resnik-based measures when using the Web as a corpus (with correlation values ranging from 0.37 to 0.48 in comparison to the 0.35–0.40 range of Resnik-based ones). Considering that they do not require any

Table 2 Similarity results and correlation factors obtained for the evaluated measures

Word pair	Resnik Human ratings	Resnik_IR	Lin_IR	Jcn_IR	Lin_ CIC$_T$_IR	JCN_ CIC$_T$_IR	PMI_IR	SCP_IC_IR	PMI_ CIC$_T$_IR	SCP_ CIC$_T$_IR
car automobile	3.9	20.929	1.773	18.135	*0.947*	*−1.54*	8.108	−7.509	*23.157*	*−1.038*
asylum madhouse	3.6	16.643	0.909	−3.348	*0.838*	*−6.396*	13.782	−9.034	*28.15*	*−1.214*
midday noon	3.6	21.949	1.343	11.207	*0.951*	*−1.856*	12.261	−8.173	*25.043*	*−2.745*
gem jewel	3.5	20.591	1.418	12.164	*0.914*	*−2.854*	11.037	−7.044	*27.701*	*−1.751*
journey voyage	3.5	20.791	1.474	13.347	*0.926*	*−3.331*	10.211	−7.866	*21.087*	*−1.223*
boy lad	3.5	15.617	1.191	5.016	*0.809*	*−6.974*	7.751	−10.78	*26.681*	*−3.762*
coast shore	3.5	16.472	1.22	5.908	*0.938*	*−1.911*	10.737	−5.563	*23.776*	*−1.512*
magician wizard	3.5	24.565	1.594	18.303	*0.957*	*−1.494*	12.513	−5.772	*25.57*	*−2.312*
tool implement	3.4	15.094	1.209	5.162	*0.89*	*−3.308*	9.216	−6.522	*23.454*	*−5.567*
furnace stove	2.6	20.64	1.238	7.954	*0.801*	*−10.232*	12.919	−7.492	*21.396*	*−3.09*
brother monk	2.4	18.827	1.285	8.329	*0.939*	*−2.125*	10.805	−7.715	*20*	*−3.501*
bird cock	2.2	22.052	1.83	20.01	*0.887*	*−5.251*	5.391	−13.316	*25.041*	*−3.889*
food fruit	2.1	15.204	1.246	6.015	*0.809*	*−7.014*	9.629	−5.146	*29.188*	*−1.488*
bird crane	2.1	22.052	1.505	14.807	*0.908*	*−4.457*	10.78	−7.746	*24.925*	*−2.173*
lad brother	1.2	15.617	1.104	2.944	*0.797*	*−7.534*	9.444	−9.431	*18.494*	*−7.29*
food rooster	1.1	15.089	1.102	2.79	*0.734*	*−10.794*	7.945	−11.49	*24.883*	*−3.137*
monk oracle	0.8	15.617	1.01	0.314	*0.743*	*−10.464*	10.06	−10.817	*23.322*	*−4.578*
coast hill	0.7	16.485	1.34	8.339	*0.818*	*−6.605*	9.38	−5.845	*23.907*	*−5.988*
journey car	0.7	14.878	1.271	6.321	*0.806*	*−7.154*	7.674	−8.119	*16.027*	*−7.854*
monk slave	0.7	15.617	1.062	1.831	*0.757*	*−9.713*	9.138	−11.155	*18.422*	*−1.371*
lad wizard	0.7	15.631	1.063	1.852	*0.753*	*−9.897*	7.934	−13.545	*18.451*	*−6.495*
forest graveyard	0.6	14.883	0.973	−0.803	*0.715*	*−11.835*	10.564	−9.403	*24.115*	*−4.51*
coast forest	0.6	17.657	1.366	9.462	*0.835*	*−6.979*	10.04	−5.77	*23.988*	*−2.905*
crane implement	0.3	22.052	1.511	14.931	*0.952*	*−2.214*	8.447	−12.272	*22.513*	*−4.662*
chord smile	0.1	17.657	1.144	4.458	*0.786*	*−9.623*	10.761	−9.339	*20.032*	*−5.214*
glass magician	0.1	18.621	1.273	7.981	*0.714*	*−11.931*	9.349	−10.554	*28.355*	*−5.853*
noon string	0.0	17.657	1.288	7.909	*0.859*	*−5.788*	8.533	−10.304	*17.127*	*−6.826*
rooster voyage	0.0	14.878	0.952	−1.47	*0.667*	*−14.846*	8.285	−14.642	*18.326*	*−8.086*
Correlation	0.884	0.403	0.357	0.364	0.665	0.678	0.376	0.486	0.455	0.739

background taxonomy, they are an effective unsupervised way to assess concept's relatedness (Turney 2001; Etzioni et al. 2005; Downey et al. 2007). Under the same conditions (rewritten in their IC-based versions), SCP outperforms PMI by a considerable margin (0.48 vs. 0.37), as it has been observed in previous works (Downey et al. 2007).

- Introducing the contextualized taxonomy-based IC computation to collocation measures, we observe clear improvements (0.45 vs. 0.37 and 0.73 vs. 0.48). As stated in Section 5, the added knowledge biases the corpus statistical analysis towards the correct word sense and guides the occurrence analysis to the taxonomic side. As expected, SCP-based function outperforms again its PMI counterpart (0.73 vs. 0.45), which cannot compete against the $\mathrm{CIC_T}$ versions of Resnik-based measures. Analyzing the numbers, $\mathrm{SCP\_CIC_T\_IR}$ is able to improve its basic version by a 50% (0.739 vs. 0.48). This is an expected improvement obtained at the cost of requiring a background taxonomy. In a more fair comparison (as both measures exploit a taxonomy), the benefit against the contextualized version of Lin and Jiang & Conrath (0.73 vs. 0.67) is more subtle (around a 9%) as, in general, they are based in the same premise. However, SCP does not include the ambiguous estimation of LCS's IC, which results in a more accurate assessment. At the end, $\mathrm{SCP\_CIC_T\_IR}$ have been able to obtain a correlation which is only a 7% worse than the original Resnik-based measures (0.739 vs. 0.794) applied to the limited tagged data of SemCor. This shows the reliability of Web-based statistics when language ambiguity is tackled.

7 Conclusions

In this paper, it has been studied the behavior of several classical semantic similarity estimation paradigms. On the one hand, IC-based approaches that exploit taxonomical knowledge are able to provide high quality results if the concept probability estimation is accurate. In classical approaches, the accuracy of this estimation is based on the preprocessing of the corpus used as background. This introduces problems about data-sparseness, due to the limited coverage of manually WordNet-based tagged data, and hampers their applicability to broader corpus or more specific ontologies.

Applying them over a massive unprocessed corpus like the Web resulted in very inaccurate similarity assessments, as absolute word occurrences provided a poor estimation of concept probabilities.

In order to minimize these problems and to maintain the scalability of the approach, we have proposed a contextualized version of IC computation which seeks for explicit word co-occurrences between evaluated concepts and their LCS in the Web. Although this is subestimation as only word observations which minimized ambiguity are taken into consideration in the concept's probability assessment, it has provided more accurate results. This shows, on one hand, that even reducing the size of the corpus, the Web provides enough resources to extract reliable conclusions. On the other hand, the calculated probabilities, even subestimated, lead to better similarity assessments due to the minimized ambiguity. It is important to note that this approach is able to provide taxonomically coherent IC estimations with a constant -low- number of web queries for non-polysemic ontologies. Resnik-like approaches would require and exponential amount according to concept's branching factor of specializations, hampering the scalability of the approach. For polysemic cases, the number of queries is linear to the number of LCS available for the pair of evaluated concepts.

At the end, applying this approach to classical measures has shown a very considerable improvement for Lin and Jiang & Conrath functions.

On the other hand, we have analyzed collocation measures which, unlike Resnik, exploit explicit term co-occurrences as an estimation of similarity. They are much more general approaches as no previous knowledge is needed and no dependency about the corpus preprocessing is introduced. So, they provide better results when estimating concept probabilities from absolute word occurrences in the Web. However, they are affected by ambiguity and the lack of semantics in assessing the type of the relation implicit in the co-occurrence observation.

Using the proposed contextualized-IC computation and extending it to explicit co-occurrence of concepts, we have also modified the basic collocation-based functions, at the cost of losing their unsupervised nature. Opposite to Resnik-like functions, in this manner, the ambiguity of *all* web queries for IC computation is minimized due to the inclusion of the LCS as context. As mentioned in Section 5, this additional knowledge biases the web search towards the correct word sense (for polysemic words) and towards the taxonomical side of the co-occurrence.

As a result, collocation measures in the form of SCP are able to provide the best results from our tests, exploiting the additional knowledge provided by the background ontology and without depending on a preprocessed corpus.

As a main contribution, the modified versions of similarity measures proposed in this paper are able to provide results for virtually any possible concept contained in WordNet (as far as they are indexed by web search engines) or any ontology (i.e. an ontology containing domain-specific classes or even instances not considered in WordNet). Moreover, this is done in a Web-scalable manner without any kind of manual intervention or pre-processing. In consequence data sparseness problems which may appear with rare concepts are greatly minimized and the generality of the measures is improved. These benefits are accomplished maintaining the results' reliability, which are only marginally worse (around a 7% less accurate for SCP_CIC$_T$_IR) when comparing them against measures using the limited tagged data of SemCor for accurate concept probability estimation.

As ongoing and future lines of research, we have beginning to apply the proposed approach to other domains and ontologies. Biomedicine, for example, is an interesting field in which standard domain ontologies such as SNOMED-CT or MeSH are available. Preliminary results obtained in that field for a standard benchmark (Pedersen et al. 2007) using SNOMED as ontology and the Web as the corpus from which compute IC, have been presented in Sánchez et al. (2009), showing improvements in comparison to a preprocessed domain corpus. Considering the lack of constraints and the generality of our approach (as the Web potentially covers any kind of knowledge), our approach can be applied and evaluated to any domain for which a widely agreed ontology exists (e.g. chemistry, computer science, etc.).

In general, any application requiring concept semantic similarity estimation which rely in concrete ontologies modeling concepts not typically covered by classical repositories like SemCor, may improve its performance by applying the measures proposed in this paper.

In particular, we are interested in applying them in unsupervised clustering methods. In this field, semantic knowledge is usually poorly exploited, since concepts are treated as categorical values without any associated semantics. The use of a semantic-based similarity to compare the objects could improve the quality of the classifications obtained. A first attempt to include semantic-based measures has been done (Batet et al. 2008) and further studies in this line are in progress. In this sense, the contextualization of the similarity calculation permits to include the background knowledge provided by specific domain ontologies (instead of WordNet) modeling concrete domain vocabulary or even instances.

Another promising area of interest for us is the use of ad-hoc ontologies for improving anonymization techniques, which are required for privacy preserving (e.g. in statistical disclosure control in data bases). A quite common way of anonymization consists of masking the information using microaggregation methods (Domingo-Ferrer and Torra 2001). Those methods introduce some modifications to the data mining clustering techniques in order to ensure the anonymity property of the individuals (Sweeney 2002). However, other methods based on non-perturbative (ontology-based) approaches could also apply the similarity measures proposed in this paper.

Finally, another interesting research line consists on extending the approach to the computation of semantic *relatedness* between concepts. As stated in the introduction, relatedness computation is a more general measure than similarity. It considers, in addition to taxonomic relations, other inter-concept non-taxonomic relationships. Even though non-taxonomic knowledge is rarer than taxonomical, large and rich ontologies such as WordNet partially models it. As stated in Section 4, as absolute term co-occurrence in the Web covers any kind of semantic relation between concepts, our approach could be easily adapted by including, as context, other types of common ontological ancestors, exploiting non-taxonomic relationships such as meronymy, holonomy, antonymy, etc.

Acknowledgements This research has been partially supported by the Spanish Government within projects ARES (CONSOLIDER-INGENIO 2010 CSD2007-00004) and E-AEGIS (TSI2007-65406-C03-02). The work is partially supported by the Universitat Rovira i Virgili (2009AIRE-04) and the DAMASK project (Data mining algorithms with semantic knowledge, TIN2009-11005). Montserrat Batet is also supported by a research grant provided by the Universitat Rovira i Virgili.

References

Batet, M., Valls, A., & Gibert, K. (2008). Improving classical clustering with ontologies. In *Proceedings of the 4th world conference of the international association for statistical computing* (pp. 137–146). Yokohama, Japan.

Berners-lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 34–43.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). WebSim: A web-based semantic similarity measure. In *Proceedings of the 21st annual conference of the Japanese society for artificial intelligence*. Miyazaki.

Brill, E. (2003). Processing natural language without natural language processing. In *Proceedings of the 4th international conference on computational linguistics and intelligent text processing* (pp. 360–369). Mexico City, Mexico.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics, 32*(1), 13–47.

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In *Proceedings of lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). New Jersey, USA.

Cilibrasi, R., & Vitanyi, P. M. B. (2006). The Google similarity distance. *IEEE Transaction on Knowledge and Data Engineering, 19*(3), 370–383.

Cimiano, P. (2006). *Ontology learning and population from text. Algorithms, evaluation and applications*. Berlin: Springer.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., et al. (2004). Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM conference on information and knowledge management* (pp. 652–659). New York: ACM.

Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. Lane, J. Theeuwes, & L. Zayatz (Eds.), *Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies* (pp. 111–134). Amsterdam: Elsevier.

Downey, D., Broadhead, M., & Etzioni, O. (2007). Locating complex named entities in Web text. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 2733–2739).

Dujmovic, J., & Bai, H. (2006). Evaluation and comparison of search engines using the LSP method. *Computer Science and Information Systems, 3*(2), 711–722.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., et al. (2005). Unsupervised named-entity extraction form the Web: An experimental study. *Artificial Intelligence, 165*, 91–134.

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. Berlin: Springer.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Ferreira da Silva, J., & Lopes, G. P. (1999). Local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of sixth meeting on mathematics of language* (pp. 369–381).

Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). *Ontological engineering* (2nd printing). Berlin: Springer.

Guarino, N. (1998). Formal ontology in information systems. In N. Guarino (Ed.), *1st international conference on formal ontology in information systems* (pp. 3–15). Trento: IOS Press.

Hotho, A., Maedche, A., & Staab, S. (2002). Ontology-based text document clustering. *Künstliche Intelligenz, 4*, 48–54.

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the international conference on research in computational linguistics* (pp. 19–33), Japan.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

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

Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation, 49*(2), 188–207.

Lemaire, B., & Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarities. *Current Psychology Letters, 18*(1). http://cpl.revues.org/document471.html. Accessed 26 May 2009.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conf. on machine learning* (pp. 296–304). San Francisco: Kaufmann.

Miller, G., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of ARPA workshop on human language technology* (pp. 303–308). Morristown: Association for Computational Linguistics.

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

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the conference of the European association for computational linguistics* (pp. 1–8). Trento, Italy.

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

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*(1), 17–30.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of 14th international joint conference on artificial intelligence* (pp. 448–453).

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

Ruch, P., Baud, R. H., Rassinoux, A. M., Bouillon, P., & Robert, G. (2000). Medical document anonymization with a semantic lexicon. In *Proceeding of the American medical informatics association symposium* (pp. 729–733).

Sánchez, D. (2008). *Domain ontology learning from the web*. Saabrucken: VDM Verlag.

Sánchez, D., Batet, M., & Valls, A. (2009). Computing knowledge-based semantic similarity from the Web: An application to the biomedical domain. In *Proceedings of the 3rd international conference on knowledge science, engineering and management* (in press).

Spence, D. P., & Owens, K. C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research, 19*, 317–330.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10*(5), 557–570.

Tadepalli, S., Sinha, A. K., & Ramakrishnan, N. (2004). Ontology driven data mining for geosciences. *Abstracts with Programs — Geological Society of America, 36*(5), 149.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth European conference on machine learning* (pp. 491–499). Freiburg, Germany.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the association for computational linguistics* (pp. 133–138). New Mexico, USA.

Yarowsky, D. (1995). Unsupervised word-sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd annual meeting of the association for computational linguistics* (pp. 189–196). Cambridge, MA.