Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content

Montserrat Batet[a1], David Sánchez[b]

[a]*CYBERCAT-Center for Cybersecurity Research of Catalonia, Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, 08860 Castelldefels (Barcelona), Spain*

[b]*CYBERCAT-Center for Cybersecurity Research of Catalonia, UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain*

Abstract

Semantic similarity measures based on the estimation of the Information Content (IC) of concepts are currently regarded as the state of the art. Calculating the IC in an intrinsic (i.e., ontology-based) way is particularly convenient due to its accuracy and lack of dependency on annotated corpora. Intrinsic IC calculation models estimate concept probabilities from the taxonomic knowledge (i.e., number of hyponyms and/or hypernyms of the concepts) modelled in an ontology. In this paper, we aim to improve the intrinsic calculation of the IC by leveraging not only the hyponyms and hypernyms of concepts, but also the explicit evidences of synonymy and polysemy that ontologies such as WordNet also model. Specifically, we propose a more accurate intrinsic estimation of the concepts' probabilities in which the IC calculation relies. We evaluate the accuracy of our proposal through a set of comprehensive experiments in which our IC calculation model is tested on a variety of IC-based similarity measures and benchmarks. Experimental results show that our proposal obtains consistently good accuracies, which vary less across measures and benchmarks than the most prominent intrinsic IC calculation models available in the literature.

Keywords: Information Content, ontology-based semantic similarity, synonymy, polysemy, WordNet.

[1] Corresponding author. Address: Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5, Parc Mediterrani de la Tecnologia, 08860 Castelldefels (Barcelona), Spain.
Tel.: +34 977559657; Fax: +34 977 559710;
E-mail: montserrat.batet@urv.cat.

1. Introduction

Semantic similarity is a core component of computational linguistics. Given two linguistic entities (words or concepts), their semantic similarity captures the resemblance between their meaning or *semantic content*. Measures quantifying the semantic similarity aim at mimicking the cognitive process by which humans create and understand textual data. Mathematically, semantic similarity measures are binary functions that numerically quantify the strength of the semantic relationship between the compared entities; thus, they can be used as alternative to the standard arithmetical distance when comparing linguistic terms. This makes semantic measures the key to apply standard data analysis and mining algorithms to textual resources. Specifically, by relying on semantic similarity measures, textual data can be characterized and managed (e.g., compared, sorted, aggregated, etc.) in terms of the semantics they convey (Rodriguez-Garcia et al. 2017). Thanks to this, semantic similarity measures have been used in a variety of data mining tasks, such as document classification (Wang and Domeniconi 2008) or data clustering (Batet 2011). Other relevant linguistically-oriented applications of semantic similarity include information retrieval (Chan et al. 2011; Kim et al. 2017), information extraction (Vicient et al. 2013), word sense disambiguation (McInnes and Pedersen 2013), paraphrase detection (Fernando and Stevenson 2008), text summarization (Gambhir and Gupta 2017) or knowledge acquisition (Sánchez et al. 2012b).

Strictly speaking, the research community understands that *similarity* measures base their assessment on *just* the taxonomic resemblance between the compared entities, that is, the evaluation of semantic evidences is constrained to "is-a" relationships. In contrast, semantic *relatedness* measures also consider non-taxonomic relationships. Our work focuses on the former type of measures, which capture the natural way by which humans hierarchically organize knowledge (Gómez-Pérez et al. 2004). Because semantics is an inherently human feature, the semantic measures proposed in the literature exploit one or several manually-created sources of knowledge. Among these, ontologies stand out as a natural (and the most popular) choice. Ontologies offer formal definitions of the conceptualizations, properties and interrelationships of the entities of a domain (Cimiano 2006). In comparison to other knowledge sources such as textual corpora, ontologies offer explicit, unambiguous and validated knowledge. Due to these advantages, a plethora of ontology-based similarity measures have been proposed in the last twenty years (Batet and Sánchez 2014). Ontology-based measures can be classified in three main families according to the mathematical principle they use to quantify the similarity:

- *Edge-counting measures* quantify the similarity inversely to the length of the minimum path of "is-a" links connecting the two concepts in the ontology (Rada et al. 1989). Therefore, these measures only consider the taxonomic ancestors (hypernyms) of the concepts in their minimum path.
- *Feature-based measures* leverage semantic evidences of similarity (e.g., common ancestors) and of dissimilarity (e.g., non-common ancestors) between concepts. These evidences are aggregated through a similarity coefficient (Sánchez et al. 2012a). In comparison with edge-counting measures, feature-based ones consider a greater amount of semantic evidences. Specifically, feature-based measures evaluate the whole set of taxonomic ancestors of the concepts, which is a superset of those in the minimum "is-a" path when ontologies model multiple taxonomic inheritance.
- *Information Content-based measures* rely on the notion of *Information Content* (IC) of a concept as a measure of the semantics conveyed by the concept. These measures quantify the resemblance between two concepts as a function of the IC of their *most informative common ancestor* (MICA), which provides evidences of similarity, and of the IC of the individual concepts, which provides evidences of dissimilarity. Two types of IC calculation models exist: corpus-based IC models, which measure the IC of concepts as the inverse of their probability of occurrence in textual corpora (Resnik 1999), and intrinsic IC models, which derive the probability of occurrence of concepts (and, thus, their informativeness) from their number of hyponyms and/or hypernyms in an ontology. Intrinsic models assume that the less hyponyms and the more hypernyms a concept has, the less probable, the more specific and the more informative the concept is (Sánchez et al. 2011; Seco et al. 2004). IC-based measures employing intrinsic IC models consider a greater amount of ontological evidences than edge-counting and feature-based measures.

Extensive evaluations on standard benchmarks and ontologies (Adhikari et al. 2018; Harispe et al. 2014; Lastra-Díaz and García-Serrano 2015a; Lastra-Díaz and García-Serrano 2015b; Sánchez and Batet 2011; Sánchez et al. 2011) regard measures based on intrinsic IC calculation as the state of the art on semantic similarity. Experimental results show that these measures tend to provide the most accurate results (i.e., similarity assessments that best mimic human similarity ratings), followed by feature-based measures and edge-counting measures. Moreover, intrinsic IC models are more convenient than corpus-

based because the former do not require a representative and semantically annotated (i.e., disambiguated) textual corpus from which derive the concept probabilities needed for IC calculation.

1.1. Contributions and plan

The differences in the accuracy among families of similarity measures sustain the following intuition*: the greater the amount of semantic evidences is considered, the more accurate the similarity assessments will be*. This constitutes the main motivation of our work. Indeed, evidences of similarity, dissimilarity or informativeness found in an ontology are the result of conscious, explicit and consensual knowledge engineering process (Gómez-Pérez et al. 2004). Therefore, it seems reasonable that the more of these evidences we consider, the better the measures will capture the cognitive notion of semantic similarity as understood by the human experts that constructed the ontology.

As introduced above, measures employing intrinsic IC models account the greatest amount of ontological knowledge. However, widely used ontologies such as WordNet contain additional taxonomic evidences (other than hyponyms and hypernyms) that we believe can contribute to improve semantic similarity assessments. According to this, we aim to improve the calculation of the IC of concepts by proposing a more accurate intrinsic estimation of the probability of occurrence of such concepts. For this, we not only leverage the number of hyponyms and hypernyms of the concepts in the ontology (as done by related works, see Section 2), but also the explicit evidences of synonymy and polysemy that ontologies such as WordNet also model. We empirically evaluate the benefits brought by our approach to the estimation of semantic similarity through a set of comprehensive experiments on a variety of IC-based similarity measures and benchmarks. Experimental results show that our proposal obtains consistently good similarity accuracies, which vary less among measures and benchmarks than the most prominent intrinsic IC calculation models available in the literature.

The rest of the paper is organized as follows. Section 2 surveys and discusses IC calculation models. Section 3 presents our approach to intrinsically calculate the IC of concepts from taxonomic knowledge. Section 4 details the evaluation testbed and reports and discusses the results of the empirical experiments. Section 5 provides the conclusions and depicts several lines of future research.

2. IC calculation models

As introduced above, IC calculation models can be classified in two types according to the knowledge source they exploit: corpus-based IC models, which rely on distribution of words in corpora, and intrinsic IC models, which rely on the semantic relationships modelled in ontologies.

In his seminal work, Resnik proposed a corpus-based IC model (Resnik 1999) that relies on the idea that infrequent terms should be more informative than common ones. Following this idea, he computed the IC of a concept c as the inverse of the probability of appearance of c in a corpus.

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

A fundamental requirement of IC-based similarity assessments is that IC values should monotonically increase as concepts are specialized (i.e., $\forall c_1, c_2 \mid c_1$ *is a hyponym of* $c_2 \Rightarrow IC(c_1) \geq IC(c_2)$). Resnik's model fulfilled this requirement by computing $p(c)$ as the number of occurrences of c plus those of the hyponyms of c. The rationale of this procedure is that the occurrence of a hyponym of c (e.g., apple) conveys the semantics of c (e.g., fruit).

Seco et al. (Seco et al. 2004) were the first to propose an intrinsic IC calculation model. They considered that the IC of a concept can be estimated as a function of its number of hyponyms in an ontology, as follows:

$$IC_{seco\ et\ al}(c) = \frac{-\log\left(\frac{|hypo(c)|+1}{max_nodes}\right)}{-\log\left(\frac{1}{max_nodes}\right)} = 1 - \frac{\log(|hypo(c)|+1)}{\log(max_nodes)}, \tag{2}$$

where $|hypo(c)|$ is the number of hyponyms of the concept c and *max_nodes* is the total number of concepts in the ontology. Note that the +1 in the numerator is added to consider the concept itself as

hyponym in order to avoid *log*(*0*) when *c* is a leaf, and *max_nodes* matches with *hypo*(*root*)+1. The denominator normalizes IC values in the range [0..1].

Corpus-based IC calculation and Seco et al.'s model rely on a similar principle. Whilst corpus-based computes $p(c)$ from the occurrence of words in a corpus, Seco et al. estimate such probability from the taxonomic structure of the ontology. Specifically, Seco et al. consider that a concept *c* is implicitly referred in text through its hyponyms (as in Resnik's calculation of $p(c)$). As a result, concepts with many hyponyms are likely to appear more probably and, therefore, will provide less information than concepts at the leaves of the taxonomy, for which the information is maximum. Numerically, *c*'s occurrence is estimated as $hypo(c)$ and $p(c)$ is estimated as the ratio between $hypo(c)$ and *max_nodes*.

Intrinsic IC models circumvent some problems related to the use of corpora. On the one hand, the accuracy of corpus-based IC assessments depends on the availability of an appropriate corpus; that is, a corpus with a representative distribution of the concepts modelled in the ontology. Otherwise, we can suffer from bias or data sparseness. On the other hand, corpus-based IC calculation requires disambiguating and (manually) annotating the words in the corpus to concepts in the ontology, so that we can compute concept probabilities in an accurate way. This is a time-consuming process that must be updated if either the ontology or the corpus changes. Intrinsic IC models are free from these handicaps and compete in accuracy with corpus-based approaches (Lastra-Díaz and García-Serrano 2015a).

The main limitation of Seco et al.'s approach is that two concepts with the same number of hyponyms but located at different depths in the taxonomic tree result in the same IC value (Zhou et al. 2008). More recent approaches have tackled this issue by leveraging additional taxonomic features.

In Sánchez et al. (Sánchez et al. 2011) the authors argued that the degree of detail, branching factor and granularity of the inner structure of the hyponym tree may hamper the consistency of intrinsic IC calculations (which is important when comparing IC values computed from different ontologies). To make the IC calculation less dependent of the inner structure of the hyponym tree, they only consider the leaves of the taxonomy below the concept, instead of the full set of hyponyms. Moreover, they also differentiate concepts at different depths by leveraging their number of hypernyms. The proposed model is consistent with the notion of probability, which is estimated as the ratio between the number of leaves of *c* (weighted by the number of hypernyms) and the total number of leaves in the taxonomy:

$$IC_{Sánchez\ et\ al.}(c) = -\log\left(\frac{\frac{|leaves(c)|}{|hypernyms(c)|}+1}{max_leaves+1}\right) \tag{3}$$

Meng et al. (Meng et al. 2012) proposed an intrinsic model that calculates the IC of a concept according to its relative depth and the depth of the hyponyms of *c*. The depth of each hyponym is considered so that concepts with highly specialized hyponyms (i.e., hyponyms in the deepest levels of the taxonomy) produce higher IC values. The relative depth of the concept is used to weight the IC, so that deeper concepts produce higher IC values. Even though this model leverages different taxonomic features (hyponyms and their depths) to estimate the informativeness of concepts, it does not aim to estimate concept probabilities in an intrinsic way.

$$IC_{Meng\ et\ al}(c) = \frac{\log(depth(c))}{\log(max_depth)} \times \left(1 - \frac{\log\left(1+\sum_{h\in hypo(c)}\frac{1}{depth(h)}\right)}{\log(max_nodes)}\right) \tag{4}$$

Adhikari et al. (Adhikari et al. 2015) extended Meng et al.'s model by also considering the ratio between the normalized number of leaves and the number of hypernyms of the concept.

$$IC_{Adhikari\ et\ al}(c) = \frac{\log(depth(c)+1)}{\log(max_depth+1)} \times \left(1 - \log\left(\frac{\frac{|leaves(c)|\times|dhyper(c)|}{max_leaves}}{hypernyms(c)}+1\right)\right) \times \left(1 - \frac{\log\left(1+\sum_{h\in hypo(c)}\frac{1}{depth(h)}\right)}{\log(max_nodes)}\right), \tag{5}$$

where $|dhyper(c)|$ is the number of hypernyms that are directly connected to *c* by a single link in the ontology.

Yuan et al. (Yuan et al. 2013) calculates the IC of a concept through the aggregation of its relative depth, its relative number of leaves and its relative number of hyponyms. Like the previous model, Yuan

et al.'s approach is not probabilistically consistent: the proposed score is an ad-hoc aggregation of taxonomic features that are positively correlated with the specificity (or informativeness) of the concept.

$$IC_{Yuan\ et\ al}(c) = f_{depth}(c)(1 - f_{leaves}(c)) + f_{hyper}(c)$$
$$f_{depth}(c) = \frac{\log(depth(c))}{\log(max_depth)}$$
$$f_{leaves}(c) = \frac{\log(|leaves(c)| + 1)}{\log(max_leaves + 1)} \quad (6)$$
$$f_{hyper}(c) = \frac{\log(|hyper(c)| + 1)}{\log(max_nodes)}$$

We can find other intrinsic models in the literature, such as of Zhou et al.'s (Zhou et al. 2008), Blanchard et al.'s (Blanchard et al. 2008), Sebti and Barfroush's (Sebti and Barfroush 2008), Sánchez and Batet's (Sánchez and Batet 2012) and Hadj Tieb et al.'s (Hadj-Taieb et al. 2014). However, recent studies have shown that they are not competitive in terms of accuracy, despite some of them being more recent than the above detailed ones (Lastra-Díaz and García-Serrano 2015a).

3. Accurate IC calculation through taxonomic knowledge

As introduced in Section 2, the inceptive idea behind intrinsic IC calculation was to derive concept probabilities, $p(c)$, from the taxonomic knowledge modelled in an ontology. For this, Seco et al. (Seco et al. 2004) assumed that concepts with many hyponyms will occur more probably than those with few hyponyms because the former can be implicitly referred in text by means of all their hyponyms. The core of the Seco et al.'s measure ($(hypo(c)+1)/max_nodes$, see numerator of eq. (2)) specifically captures this notion in a probabilistically consistent manner: the numerator provides an estimation of the number of occurrences of the concept, whereas the denominator accounts the total number of occurrences of all the concepts in the ontology (*max_nodes*), which corresponds to those of the root node. The estimated probability provides values in the (0..1] range, which are also independent of the size of the ontology.

Several refinements of Seco et al.'s measure, such as Sánchez et al. (Sánchez et al. 2011) , also follow the same probabilistically consistent design. In contrast, more recent models, such as Meng et al. (Meng et al. 2012), Adhikari et al. (Adhikari et al. 2015), Yuan et al. (Yuan et al. 2013), Hadj Tieb et al. (Hadj-Taieb et al. 2014) or Zhou et al. (Zhou et al. 2008), propose ad-hoc scores based on the aggregations of several taxonomic features (hypernyms, hyponyms/leaves and/or depth). Even though these features are aggregated so that they positively correlate with the specificity (or informativeness) of concepts, these models are less intuitive and generalizable than those that are consistent with the notion of probability. In fact, due to their lack of mathematical consistency, they estimate the specificity of concepts in arbitrary units, rather than in *units of information* (as the Resnik's information theoretic approach does (Resnik 1995)).

In this work, we go back to the roots of intrinsic IC calculation. Specifically, we aim to improve the intrinsic assessment of concept probabilities by leveraging as much taxonomic knowledge as available in the ontology while, at the same time, maintaining the mathematical consistency of the calculations. To achieve this goal, we analysed the taxonomic knowledge that can be found in large and widely used ontologies such as WordNet (and also in domain-specific ontologies, such as SNOMED-CT). From this analysis we derived the following insights on the relationship between the taxonomic knowledge found in ontologies and the probability of occurrence of the concept c to which it refers:

1. *hyponyms*(c): the number of hyponyms of a concept provides the most important insight on the probability of occurrence of a concept. The semantics encompassed by general concepts are recursively specialized in ontologies by means of taxonomic specializations (i.e., hyponyms). Therefore, as argued by Seco et al. (Seco et al. 2004) we can assume that concepts with a large number of hyponyms will occur more probably because such concepts can be implicitly referred by means of all of their specializations. As show in Definition 1, *hyponyms*(c) should also include the concept c itself because, obviously, c represents the most explicit way to refer to c; in this manner, if c is a leaf node, that is, a concept at the bottom of the taxonomic tree, then $|hyponyms(c)|=1$.

Definition 1. Let the concept subsumption ($\leq$) be a binary relation $\leq : C \times C$, being C the set of concepts in the ontology, where $c_1 \leq c_2$ means that c_1 is a hierarchical specialization of c_2. The set of *hyponyms* of a concept c is defined as:

$$hyponyms(c) = \{c_i \in C \mid c_i \leq c\} \cup \{c\}$$

Some of the approaches discussed in Section 2 (Sánchez et al. 2011; Yuan et al. 2013) use the *leaves*(c) of the concept, that is, the hyponyms at the bottom of the taxonomy, instead of *hyponyms*(c). Even though *leaves(c)* and *hyponyms(c)* provide similar evidences, the former omits the inner concepts of the taxonomy (e.g., accounting around 21% of the total number of concepts in WordNet, see Table 1).

2. *hypernyms*(c): hypernyms represent the inverse to hyponyms: the more hypernyms a concept has, the more specific it is because its semantics has been iteratively refined through a number of specializations. Therefore, concepts with many hyponyms will be less probable than those with few. Even though the evidences provided by the set of hypernyms of a concept may seem redundant to those brought by its hyponyms, considering the former helps to differentiate concepts with the same number of hyponyms but different specificities (Sánchez et al. 2011). Concretely, hypernyms are useful to differentiate the specificity of leaf concepts, which have no hyponyms other than themselves. This is important because leaves account most of the concepts modelled in ontologies (e.g., around a 79% in WordNet, see Table 1). Finally, the set of hypernyms of a concept grows in two cases: as the concept becomes specialized and in case of multiple inheritance. We believe that both cases provide evidences of specificity and should be equally considered to estimate probabilities. Therefore, given two concepts at the same "depth" in the taxonomic tree, the concept with multiple inheritance (which will be the one with the largest set of hypernyms) is more specific because the semantics it conveys is the intersection of the semantics conveyed by all its immediate ancestors. According to Definition 2, *hypernyms*(c) also consider c, so that $|hypernyms(root)|=1$.

Definition 2. Let the concept subsumption ($\leq$) be a binary relation $\leq : C \times C$, being C the set of concepts in the ontology. The set of *hypernyms* of a concept c is defined as:

$$hypernyms(c) = \{c_i \in C \mid c_i \geq c\} \cup \{c\}$$

depth(c), which is the position of the concept in the taxonomy relative to the root node, is an alternative to *hypernyms*(c) that has been also considered in some intrinsic IC models (see Section 2). *depth(c)* is computed as minimum path between the root node of the ontology and c; +1 is added to account concepts rather than number of links. According to Definition 3, *depth*(c)=|*hypernyms*(c)| if no multiple inheritance is considered.

Definition 3. The *depth* of a concept c is defined as:

$$depth(c) = min_path(root, c) + 1$$

3. *synonyms*(c): many ontologies model the set of words with equivalent meanings that can be used to refer to a concept c in corpora. In WordNet, the set of synonyms of a concept is called *synset*. Other ontologies, such as SNOMED-CT, also include synonyms, lexicalizations, acronyms and other interchangeable words with equivalent meanings for each concept. Synonyms are the result of the evolution of human languages and are used as a linguistic resource to minimize redundancy when a concept occurs multiple times. Therefore, general concepts of more common (probable) use tend to have more synonyms than rarer more specific ones. Even though synonyms have been neglected by the intrinsic IC calculation models proposed in the literature, due to the former reason, we believe they can positively contribute to the estimation of probabilities. Moreover, even though perfectly interchangeable synonyms exist, it is more common that synonyms present subtle nuances among their meanings and, therefore, they could be considered as different (but closely related) concepts that, because of their similarity, have not been strictly differentiated (e.g., as hyponyms) in the knowledge base. In this latter case, the synonyms of a concept can complement the evidences provided by the hyponyms. In fact, concept synonyms are finer grained than concept hyponyms: whereas WordNet encompasses 81,426 concepts, it models 145,104 synonym-concept pairs, which is an amount 78% larger than the number of concepts (see Table 1). Notice that, according to Definition 4, $|synonyms(c)| \geq 1$ for all c.

Definition 4. Let synonymy (≡) be a binary relation ≡ : $L \times C$, being C the set of concepts in the ontology and L the set of words (or lemmas) in the ontology, where $l \equiv c$ means that l is a word or lemma that conveys the meaning of c. The set of *synonyms* of a concept c is defined as:

$$synonyms(c) = \{l_i \in L \mid l_i \equiv c\}$$

4. *polysemy*(l): even though concepts are not ambiguous, the words or lemmas l used to refer to them in a context are. These words are in fact the synonyms in the *synset* associated to the concept, and each word may be or not polysemic, or may have different levels of polysemy. Ontologies such as WordNet explicitly model polysemy by linking words to all their possible senses, that is, the concepts to which each word may refer to. Like synonymy, polysemy is a common phenomenon affecting more than 19% of the words modelled in WordNet and, for those, the average polysemy count is around 2.77 senses per word (see Table 1). Even though polysemy has been also neglected by intrinsic IC calculation models, we also believe that it may have a contribution to the estimation of concepts' probabilities. Specifically, when the synonyms of a concept are considered to estimate its probability, the contribution of each synonym may be weighted by their polysemy, because the occurrences of words with different meanings should be split among those meanings (each one corresponding to a different concept). In this way, highly polysemic synonyms of a concept would contribute less to the estimation of the concept's probability than non-polysemic ones. Notice that, according to Definition 5, $|polysemy(l)|=1$ for non-polysemic words.

Definition 5. Being C the set of concepts in the ontology and L the set of words (or lemmas) in the ontology, the polysemy of a word l is defined as follows:

$$polysemy(l) = \{c_i \in C \mid l \equiv c_i\}$$

By relying on the former analysis, we present now the design of our intrinsic IC calculation model, which progressively refines the estimation of probabilities by incorporating as much taxonomic knowledge as available in current ontologies. At its core, our model intrinsically estimates the occurrence of a concept c (which we called *iOccur*(c)) according to its number of hyponyms, as proposed by Seco et al. and as discussed in point 1) above; that is, the occurrence of a concept c is directly proportional to its number of hyponyms, because the latter represent the means of referring to the concept in a discourse.

$$iOccur(c) = \sum_{\forall h_i \in hyponyms(c)} 1 \tag{7}$$

Notice that $iOccur(c) \geq 1$ for all c because, as detailed in point 1) above, $hyponyms(c) \geq 1$.

Then, as discussed in point 3) above, we consider that concepts with many synonyms tend to occur more frequently, because this greater occurrence is the very reason for creating synonyms in human languages. Also, due to the same argument as above, the synonyms of a concept c represent alternative ways of referring to that concept. In our model, *synonyms*(c) are leveraged for each of the hyponyms of c, that is, they are used to weight the contribution of each h_i in *hyponyms*(c) in the assessment of c's occurrences, as follows:

$$iOccur(c) = \sum_{\forall h_i \in hyponyms(c)} \sum_{\forall s_k \in synonyms(h_i)} 1 \tag{8}$$

With the above formulation, concepts with the same number of hyponyms and synonyms are indistinguishable. This commonly happens for the leaves of the taxonomy, for which *hyponyms*(c)=1 (i.e., the leaf itself) and, usually, *synonyms*(c)=1 (due to the high specificity of c). To be able to distinguish these concepts according to the evidences of specificity, we incorporate their hypernyms according to the argumentation in point 2) above. Concretely, we weight each h_i in *hyponyms*(c) (many of which would be indistinguishable leaf nodes), by their *hypernyms*(h_i), as follows:

$$iOccur(c) = \sum_{\forall h_i \in hyponyms(c)} \frac{\sum_{\forall s_k \in synonyms(h_i)} 1}{\sum_{\forall H_j \in hypernyms(h_i)} 1} \quad (9)$$

So far, we have given the same weight to all the synonyms (in the numerator) and all the hypernyms (in the denominator). In the following, we further refine their contribution in the assessment of *c*'s occurrences by leveraging additional taxonomic evidences. For the hypernyms in the denominator, we weight their contribution by their relative position in the taxonomic tree, that is, their *depth* (see point 2) above); specifically, the contribution of each *hypernym*(h_i) would be proportionally smaller as they become more specific (i.e., "deeper") in the ontology. This is consistent with studies highlighting that the semantic differences between concepts tend to diminish as they become more specialized because, usually, the criteria that best differentiates the meaning of concepts is used first during the specialization process (Wu and Palmer 1994).

$$iOccur(c) = \sum_{\forall h_i \in hyponyms(c)} \frac{\sum_{\forall s_k \in synonyms(h_i)} 1}{\sum_{\forall H_j \in hypernyms(h_i)} depth(H_j)} \quad (10)$$

Regarding the synonyms in the numerator of eq (9), according to the arguments in point 4) above, we inversely weight their contribution by their polysemy; that is, the number of senses of the synonym. In this way, polysemic synonyms would contribute less toward the estimation of the occurrences of *c*, because the number of occurrences of such synonyms should be split among all their meanings (and this does not happen for non-polysemic synonyms). With this, the final formulation of our intrinsic IC calculation model is as follows:

$$iOccur(c) = \sum_{\forall h_i \in hyponyms(c)} \frac{\sum_{\forall s_k \in synonyms(h_i)} \frac{1}{|polysemy(s_k)|}}{\sum_{\forall H_j \in hypernyms(h_i)} depth(H_j)} \quad (11)$$

As discussed above, $\{|hyponyms(c)|, |synonyms(h_i)|, |hypernyms(h_i)|, depth(H_j)$ *and* $|polysemy(s_k)|\} \geq 1$, which ensures that no indeterminations happen during the calculation of eq. (11) and, also, that eq. (11)≥ 0 for all *c*.

According to eq. (11), the contribution of the synonyms of each hyponym in the concept probability is a function of the polysemy of the synonym. If the synonyms are non-polysemic or all of them have the same level of polysemy, then they will have identical contribution towards the concept probability. As a counterpart, if synonyms have different level of polysemy, they will have different contribution. Hyponyms with equally informative synonyms sets are distinguished by their relative depth in the ontology, as shown in the denominator: deeper hyponyms contribute less to the concept's probability than shallower ones.

Finally, *iOccur*(*c*) can be normalized by the occurrence of the *root* concept, which is the most general concept of the ontology; that is *iOccur*(*root*). In this way, we obtain an estimation of the probability of the *c* in the standard range (0..1], and this estimation is independent of the size and granularity of the ontology. This estimated probability can then be used to intrinsically asses the information content of *c* (*iIC*(*c*)), as follows:

$$iIC(c) = -\log \frac{iOccur(c)}{iOccur(root)}, \quad (12)$$

which produces IC values in the range [0..+∞), as in Resnik's model.

Table 1. Topological statistics of the taxonomy of WordNet (version 2.1).

Feature	Quantity
Concepts	81,426
Synonyms	117,097
Synonym-concept pairs	145,104
Polysemic synonyms	15,776 (19.3%)
Average polysemy (for polysemic synonyms)	2.77
Maximum polysemy	33
Inner concepts	17,020 (21%)
Leaf concepts	64,406 (79%)
Concepts with multiple inheritance	2,128 (2.6%)
Maximum depth	18

As stated in Section 2, to be mathematically consistent, IC-based similarity measures require that IC values monotonically increase as one moves down in the taxonomy of concepts; that is, that the specializations are more informative than all their ancestors. We next prove that our IC calculation model fulfils this property.

Proposition 1. The function of eq. (12) monotonically increases as concepts are specialized in the ontology.

Proof.

$$\forall\, c_1, c_2 \in C \mid c_1 \leq c_2 \Rightarrow hyponyms(c_1) \subset hyponyms(c_2) \Rightarrow$$

$$\Rightarrow \sum_{\forall h_i \in hyponyms(c_1)} \frac{\sum_{\forall s_k \in synonyms(h_i)} \frac{1}{|polysemy(s_k)|}}{\sum_{\forall H_j \in hypernyms(h_i)} depth(H_j)} < \sum_{\forall h_i \in hyponyms(c_2)} \frac{\sum_{\forall s_k \in synonyms(h_i)} \frac{1}{|polysemy(s_k)|}}{\sum_{\forall H_j \in hypernyms(h_i)} depth(H_j)} \Rightarrow$$

$$\Rightarrow iOccur(c_1) < iOccur(c_2) \Rightarrow iIC(c_1) > iIC(c_2)\ \square$$

4. Evaluation

In this section, we evaluate the accuracy of our IC calculation model and compare it with the state-of-the-art. Because improving the accuracy of semantic similarity assessments is our final goal, the evaluation is based on the performance of the IC calculation in IC-based similarity measures. The standard way to evaluate similarity assessments is to calculate the Pearson's correlation (eq. (13)) among the similarity values obtained by a certain similarity measure (X) and those provided by human experts (Y) over a set of term pairs. A correlation value of 1 (-1) describes a perfect (inverse) relationship between the computerized results and the human ratings (which is the ultimate goal of similarity measures), whereas values near 0 correspond to random assessments.

$$Pearson_correlation(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{13}$$

For evaluation purposes, we employed the most representative and well-stablished IC-based similarity measures available in the literature: Resnik's (Resnik 1995), Lin's (Lin 1998) and Jiang & Conrath's (Jiang and Conrath 1997). Even though more recent measures exist, such as (Adhikari et al. 2018; Lastra-Díaz and García-Serrano 2015b; Meng and Gu 2012; Pirró 2009), recent studies have shown that they do not significantly improve Lin's and/or Jiang & Conrath's measures in WordNet (Lastra-Díaz and García-Serrano 2015b). Moreover, many of them can be expressed as parametrizations of the similarity coefficients in which Lin's and Jiang & Conrath's measures rely (Harispe et al. 2014; Sánchez and Batet 2011), thus leading to similar or even equivalent results.

Resnik was the first to employ the notion of IC of concepts to measure their similarity. He stated that the similarity of two concepts $<c_1, c_2>$ depends on the amount of information they share. This shared information is represented by the IC of the most specific common ancestor (MICA) that subsumes the two concepts in an ontology (Resnik 1995):

$$sim_{res}(c_1, c_2) = IC(MICA(c_1, c_2)) \quad (14)$$

One of the limitations of Resnik's metric is that any pair of concepts with the same MICA will be considered equally similar. To tackle this problem, both Lin (Lin 1998) and Jiang & Conrath (Jiang and Conrath 1997) improved Resnik's measure by also considering the IC of each individual concept $<c_1, c_2>$ as indication of dissimilarity.

Lin's measure calculates the similarity as the ratio between the information the two concepts have in common (represented by the IC of their MICA) and the information conveyed by each of them (Lin 1998). Lin's measure uses Dice's coefficient of similarity (Dice 1945):

$$sim_{lin}(c_1, c_2) = \frac{2 \times IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (15)$$

The measure proposed by Jiang & Conrath is based on quantifying the distance of the taxonomic path separating the two concepts as the difference between the IC of each concept and their MICA (Jiang and Conrath 1997):

$$dis_{j\&c}(c_1, c_2) = (IC(c_1) + IC(c_2)) - 2 \times IC(MICA(c_1, c_2)) \quad (16)$$

We used WordNet 2.1 (Fellbaum 1998) as reference ontology in our experiments. WordNet is, by far, the most widely used knowledge base in ontology-based similarity assessments (Batet and Sánchez 2014). As introduced in Section 3, WordNet contains 81,426 concepts, which are organized in *synsets* (encompassing 117,097 words).

As evaluation data, we employed a variety of widely used benchmarks of term pairs meant for semantic similarity evaluation:

- Rubenstein & Goodenough's benchmark (R&G) (Rubenstein and Goodenough 1965) consists of 65 word pairs with averaged similarity values provided by a group of 51 students in a scale from 0 (unrelated) to 4 (synonyms). The average pairwise correlation among the ratings provided by the 51 students is 0.85.
- Miller & Charles benchmark (M&C) (Miller and Charles 1991) is made up of a subset of 30 word pairs of the previous benchmark, which were re-evaluated by 38 undergraduate students in the same 0-4 scale.
- Resnik's benchmark (Resnik 1995) consists of 28 word pairs from the M&C benchmark, which were re-evaluated by 10 computer science graduate students and post-doc researchers. The average pairwise correlation among human ratings is 0.9015. Even though R&G, M&C and Resnik's benchmarks share word pairs, their similarity ratings are different and come from different types of individuals; therefore, they represent different views on semantic similarity. So, it is interesting to test all three benchmarks in order to assess the generality of the similarity calculation methods under this variety of views.
- Agirre et al.'s benchmark (Agirre et al. 2009) is composed by 203 term pairs with similarity ratings extracted from the WordSim-353 dataset. The average pairwise correlation between human ratings is 0.667. After removing verbs and other terms not found in WordNet, we obtained 197 pairs that we used in our experiments.

Because words in the benchmarks above are not disambiguated (i.e., they are not univocally associated concepts) and some of them are polysemic, similarity values may vary according to the sense considered when mapping the words to WordNet *synsets*. To tackle this issue, we computed the similarity of each pair of words as the maximum value obtained from measuring the similarity for the Cartesian product of the *synsets* (senses) to which the words belong. The rationale of this practice, which is also followed by related works (Lastra-Díaz and García-Serrano 2015a; Sánchez et al. 2011), is the fact that given two non-disambiguated words, human evaluators will tend to consider their closest meanings to evaluate their similarity.

The IC-based similarity measures introduced above have been evaluated with the four benchmarks and with five state-of-the-art IC calculation models. As introduced in Section 2, according to the comprehensive experiments reported in (Lastra-Díaz and García-Serrano 2015a) the best performing intrinsic IC models are: Seco et al. (Seco et al. 2004), Sánchez et al. (Sánchez et al. 2011), Meng et al. (Meng et al. 2012) and Yuan et al. (Yuan et al. 2013), in no particular order. Moreover, we also added the

more recent model by Adhikari et al. (Adhikari et al. 2015), which was not considered in (Lastra-Díaz and García-Serrano 2015a). The results brought by these models to the IC-based similarity assessment have been compared with those provided by our proposal.

To evaluate the individual contribution of the new taxonomic evidences considered in our model (i.e., synonymy and polysemy), we employed two versions of our proposal: the first one uses the complete assessment of concept occurrences formalized in eq. (11), which considers synonymy and polysemy, whereas the second one is a simplified version that does not consider polysemy at all, that is, eq. (10) is used instead eq. (11) in the final expression (eq. (12)).

Tables 2 to 5 report the Pearson correlation obtained by the three IC-based similarity measures with the different IC calculation models considered and the four benchmarks. Notice that the J&C measure, being a distance instead of a similarity measure, reports negative (i.e., inversely correlated) values. The *p*-value of the correlation is lower than 0.00001 in all cases, which indicates that the results and differences reported are statistically significant.

Table 2. Pearson correlation for the different IC calculation models (rows) and similarity measures (columns) with the R&G benchmark. Numbers in parentheses state the relative ranking of IC models for each similarity measure.

IC Measure / *IC model*	**Resnik**	**Lin**	**J&C**
Seco et al.	0.8405 (4th)	0.8659 (6th)	-0.8582 (6th)
Sánchez et al.	0.8470 (3rd)	0.8581 (7th)	-0.8603 (5th)
Meng et al.	0.8384 (5th)	0.8702 (3rd)	-0.8643 (2nd)
Adhikari et al.	0.8381 (6th)	**0.8709 (1st)**	-0.8642 (3rd)
Yuan et al.	0.8346 (7th)	0.8707 (2nd)	-0.8556 (7th)
This work (synonymy)	**0.8521 (1st)**	0.8682 (5th)	**-0.8649 (1st)**
This work (synonymy+polysemy)	0.8508 (2nd)	0.8685 (4th)	-0.8623 (4th)

Table 3. Pearson correlation for the different IC calculation models (rows) and similarity measures (columns) with the M&C benchmark. Numbers in parentheses state the relative ranking of IC models for each similarity measure.

IC Measure / *IC model*	**Resnik**	**Lin**	**J&C**
Seco et al.	0.8023 (7th)	0.8359 (7th)	**-0.8826 (1st)**
Sánchez et al.	0.8297 (5th)	0.8416 (6th)	-0.8680 (4th)
Meng et al.	**0.8518 (1st)**	0.8596 (2nd)	-0.8445 (6th)
Adhikari et al.	0.8515 (2nd)	**0.8602 (1st)**	-0.8450 (5th)
Yuan et al.	0.8288 (6th)	0.8450 (5th)	-0.8395 (7th)
This work (synonymy)	0.8360 (3rd)	0.8478 (4th)	-0.8773 (3rd)
This work (synonymy+polysemy)	0.8355 (4th)	0.8486 (3nd)	-0.8774 (2nd)

Table 4. Pearson correlation for the different IC calculation models (rows) and similarity measures (columns) with the Resnik benchmark. Numbers in parentheses state the relative ranking of IC models for each similarity measure.

IC Measure / *IC model*	**Resnik**	**Lin**	**J&C**
Seco et al.	0.8354 (7th)	0.8755 (6th)	-0.8947 (3rd)
Sánchez et al.	0.8503 (6th)	0.8615 (7th)	-0.8756 (4th)
Meng et al.	**0.8691 (1st)**	0.8830 (4th)	-0.8666 (6th)
Adhikari et al.	0.8690 (2nd)	0.8841 (2nd)	-0.8675 (5th)
Yuan et al.	0.8554 (5th)	0.8776 (5th)	-0.8625 (7th)
This work (synonymy)	0.8598 (3rd)	0.8835 (3rd)	-0.8952 (2nd)
This work (synonymy+polysemy)	0.8576 (4th)	**0.8871 (1st)**	**-0.8997 (1st)**

Table 5. Pearson correlation for the different IC calculation models (rows) and similarity measures (columns) with the Agirre et al. benchmark. Numbers in parentheses state the relative ranking of IC models for each similarity measure.

IC Measure / *IC model*	**Resnik**	**Lin**	**J&C**
Seco et al.	0.6599 (5th)	0.6814 (4th)	**-0.6674 (1st)**
Sánchez et al.	0.6561 (7th)	0.6628 (7th)	-0.6583 (4th)
Meng et al.	0.6681 (3rd)	0.6784 (6th)	-0.6564 (6th)
Adhikari et al.	0.6683 (2nd)	0.6797 (5th)	-0.6565 (5th)
Yuan et al.	**0.6684 (1st)**	**0.6843 (1st)**	-0.6506 (7th)
This work (synonymy)	0.6665 (4th)	0.6816 (2nd)	-0.6654 (2nd)
This work (synonymy+polysemy)	0.6583 (6th)	0.6815 (3rd)	-0.6647 (3rd)

A first analysis of the results shows that the differences among the IC calculation models are quite minor. The reason for this is the fact that most of the correlation values obtained by IC-based measures are very close (or even surpass) the average pairwise correlation among human ratings for each benchmark (i.e., 0.85 for R&G, 0.9015 for Resnik and 0.667 for Agirre et al.). Being the latter upper bounds for computerized similarity measures, the results suggest that the combination of IC-based measures with state-of-the-art IC calculation models produce results that are almost as good as they can be.

The behaviour of the IC calculation models with respect to the IC measures is also consistent: Resnik's measure provides the lowest accuracy, whereas Lin's and Jiang & Conrath measures perform very similar. The fact that the latter are able to differentiate concepts with the same MICA but with different levels of specificity (IC) contributes to improve the accuracy of the similarity assessments (Sánchez et al. 2011).

The results also show that there is not a clear winner among the IC calculation models for the different measures and benchmarks, with Meng et al., Adhikari et al., Yuan et al. and our proposal scoring the largest number of wins (best correlation values are presented in boldface). This is consistent with recent studies (Lastra-Díaz and García-Serrano 2015a), which also concluded that the best performing IC models offer similar results and that their relative performance depend on the measure and benchmark employed. To better analyse the latter, we included in the tables the relative ranking of correlation values brought by each model for each measure (see ranks in parentheses). These show a large variability for most of the models (Seco et al., Adhikari et al., Meng et al. and Yuan et al. range from 1st to 6th or 7th quite uniformly). In fact, the largest variability is observed for models like Yuan et al.'s, which use ad-hoc aggregations of ontological features. Even though the latter exploits a significant amount of taxonomic evidences (i.e., depth, leaves, hypernyms) and is able to maximize the accuracy of a particular measure in a concrete benchmark, its "overfitted" design hampers its overall performance. In contrast, our proposal (in both of its versions) presented the smallest variability, and consistently ranked among the 1st and 4th places in most cases. This indicates that, even though our model is not *absolutely better* than the state of the art for all the measures and benchmarks, it produces *consistently good* results. As anticipated in Section 3, this suggests that IC calculation models that intrinsically assess concept probabilities in a consistent manner generalize better than approaches proposing ad-hoc aggregations of ontological features.

Comparing the results of the measures that follow a probabilistically-consistent design, i.e., Seco et al.'s, Sánchez, et al.'s and ours, sustains our hypothesis that the more semantic evidences are considered in the assessment of IC, the more accurate the results will be; whereas Seco et al. only considers the hyponyms of the concepts and Sánchez et al. the hypernyms and the leaves, our approach leverages the hyponyms, synonyms (weighted by their polysemy) and hypernyms (weighted by their relative depth). As far as we know, the latter account for all the taxonomic evidences that can be found in current ontologies. Even though ontologies model other semantic relationships, such as meronymy, those are not strictly taxonomic knowledge and would contribute to assess the *relatedness* of concepts, rather than their *similarity*.

Comparing the two versions of our model we don't see a clear winner. If we carefully consider the results of all the benchmarks, the version that only considers synonymy scores better in more occasions, but with a very minor margin. One possible explanation for this phenomenon is the fact that WordNet implements polysemy in a very fine-grained way: more than 35% of nouns have more than 2 senses and some of them have more than 30. In fact, some researchers have criticized the excessive polysemy of WordNet (in comparison with the actual usage of polysemic words in society) (Clark et al. 2006) and proposed ways to coarsen it (Palmer et al. 2007). Related to this, some researchers have studied WordNet's polysemy and classified senses according to their polysemy types into metaphoric, specialization and homonymy types (Freihat et al. 2016). These types may be used to filter excessive polysemy and, as a result, may have a positive influence in the similarity assessments. We leave this possibility as future work.

In any case, given that evaluating terms' polysemy makes the IC calculation more complicated and costlier, it seems that the simplified version of our model is a more reasonable choice.

5. Conclusions

We have proposed an IC calculation model that leverages all the semantic evidences of taxonomic knowledge available in current ontologies. The design of our proposal carefully aggregates these evidences in order to be consistent with the notion of *probability of occurrence* of concepts in which the calculation of the IC is based. Our model is also mathematically consistent and fulfils the monotony

property, which is relevant in some applications (Sánchez and Batet 2017). Thanks to this mathematically consistent design, our model provided also consistently accurate results when applied to standard IC-based measures. Even though these results were *not (significantly) better* than related works in all cases, they were *consistently good* in all cases. This was not the case for other models that showed a greater variability across measures and benchmarks due to their "overfitted" (i.e., less general) designs. The fact that our model performs well in general alleviates the problem of selecting the best suited combination of IC-based measure and IC calculation model (Lastra-Díaz and García-Serrano 2015a) and makes our proposal the natural choice in a blind scenario.

Our work is also aligned with recent studies (Harispe et al. 2014), which conclude that it is better to focus on improving the basis in which semantic similarity measures rely, that is, the assessment of the commonalities, differences and specificity (informativeness) of concepts rather than on proposing new similarity measures (many of which have demonstrated to be equivalent (Harispe et al. 2014)). Moreover, the calculation of the IC of concepts is also of great interest in other tasks, such as topic extraction (Sánchez et al. 2013), user profiling (Viejo et al. 2012) or disclosure risk assessment (Viejo and Sánchez 2016).

As future work, we plan to tackle some of the limitations identified in Section 4. First, regarding the limitations of WordNet-based benchmarks, we plan to evaluate our method in other ontologies modelling synonymy and polysemy (such as SNOMED-CT) and, also, in a combination of several overlapping ontologies (Batet et al. 2014). Moreover, as discussed in the previous section, polysemy types (Freihat et al. 2016) may be used to filter out excessive polysemy, which may have a positive influence in the similarity assessments. Finally, we also plan to extend our model so that it captures semantic evidences other than taxonomic ones; with this, it would be possible to use intrinsic IC calculation models in semantic relatedness measures (Pirrò and Euzenat 2010).

Acknowledgements

This work was partly supported by the European Commission (H2020-700540 project "CANVAS"), by the Spanish Government (projects TIN2014-57364-C2-R "SmartGlacis", TIN2015-70054-REDC "Red de excelencia Consolider ARES" and TIN2016-80250-R "Sec-MCloud"). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of UNESCO.

References

Adhikari A, Dutta B, Dutta A, Mondal D, Singh S (2018) An intrinsic information content-based semantic similarity measure considering the disjoint common subsumers of concepts of an ontology Journal of the Association for Information Scence and Technology 69:1023-1034

Adhikari A, Singh S, Dutta A, Dutta B A novel information theoretic approach for finding semantic similarity in WordNet. In: TENCON 2015 IEEE Region 10 Conference, Macao, China, 2015. IEEE, pp 1-6

Agirre E, Alfonseca E, Hall K, Kravalova J, Pasca M, Soroa A A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, 2009. pp 19-27

Batet M (2011) Ontology based semantic clustering AI Communications 24:291-292

Batet M, Harispe S, Ranwez S, Sánchez D, Ranwez V (2014) An information theoretic approach to improve semantic similarity assessments across multiple ontologies Information Sciences 283:197-210

Batet M, Sánchez D (2014) Review on Semantic Similarity. In: Encyclopedia of Information Science and Technology (3rd edition). IGI Global, pp 7575-7583

Blanchard E, Harzallah M, Kuntz P A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Proceedings of 18th European Conference on Artificial Intelligence (ECAI), Patras, Greece, July 21-25 2008. IOS Press, pp 20-24

Cimiano P (2006) Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag,

Clark P, Harrison P, Jenkins T, Thompson J, Wojcik R (2006) From WordNet to a Knowledge Base. Paper presented at the AAAI 2006 Spring Symposium on Formalizing and Compiling Background Knowledge,

Chan LWC, Liu Y, Shyu CR, Benzie IFF (2011) A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity Engineering Applications of Artificial Intelligence 24:1398–1409

Dice LR (1945) Meaures of the amount of ecologic association between species Ecology 26:297-302

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

Fernando S, Stevenson M (2008) A Semantic Similarity Approach to Paraphrase Detection. Paper presented at the 11th Annual Research Colloqium Computational Linguistics UK (CLUK 2008),

Freihat AA, Giunchiglia F, Dutta B A taxonomic classification of WordNet polysemy types. In: 8th Global WordNet Conference 2016, Bucharest, Romania, 2016. pp 105-113

Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey Artificial Intelligence Review 47:1–66

Gómez-Pérez A, Fernández-López M, Corcho O (2004) Ontological Engineering. 2nd edn. Springer-Verlag,

Hadj-Taieb MA, Ben-Aouicha M, Ben-Hamadou A (2014) A new semantic relatedness measurement using WordNet features Knowledge and Information Systems 41:467-497

Harispe S, Sánchez D, Ranwez S, Janaqi S, Montmain J (2014) A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain Journal of Biomedical Informatics 49:38-53

Jiang JJ, Conrath DW Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: International Conference on Research in Computational Linguistics, ROCLING X, Taipei, Taiwan, Sep 1997. pp 19-33

Kim S, Fiorini N, Wilbur WJ, Lu Z (2017) Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents Journal of Biomedical Informatics 75:122-127

Lastra-Díaz JJ, García-Serrano A (2015a) A new family of information content models with an experiemental survey on WordNet Knowledge-Based Systems 89:509-526

Lastra-Díaz JJ, García-Serrano A (2015b) A novel family of IC-based similarity measures with a detailed experimental survey on WordNet Engineering Applications of Artificial Intelligence 46:140-153

Lin D An Information-Theoretic Definition of Similarity. In: Shavlik J (ed) Fifteenth International Conference on Machine Learning, ICML 1998, Madison, Wisconsin, USA, July 24-27 1998. Morgan Kaufmann, pp 296-304

McInnes BT, Pedersen T (2013) Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text Journal of Biomedical Informatics 46:1116-1124

Meng L, Gu J A new model for measuring word sense similarity in WordNet. In: 4th International Conference on Advanced Communication and Networking, Jeju, Korea, 2012. pp 18-23

Meng L, Gu J, Zhou Z (2012) A new model of information content based on concept's topology for measuring semantic similarity in WordNet International Journal of Grid and Distributed Computing 5:81-93

Miller GA, Charles WG (1991) Contextual correlates of semantic similarity Language and Cognitive Processes 6:1-28

Palmer M, Dang H, Fellbaum C (2007) Making fine-grained and coarse-grained sense distinctions, both manually and automatically Natural Language Engineering 13:137-163

Pirró G (2009) A semantic similarity metric combining features and intrinsic information content Data & Knowledge Engineering 68:1289-1308 doi:10.1016/j.datak.2009.06.008

Pirrò G, Euzenat J A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: International Semantic Web Conference, 2010. pp 615-630

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 doi:10.1109/21.24528

Resnik P Using information content to evalutate semantic similarity in a taxonomy. In: Mellish CS (ed) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada, 1995. Morgan Kaufmann Publishers Inc., 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 doi:10.1613/jair.514

Rodriguez-Garcia M, Batet M, Sánchez D (2017) A Semantic Framework for Noise Addition with Nominal Data Knowledge-Based Systems 122:103-118

Rubenstein H, Goodenough J (1965) Contextual correlates of synonymy Communications of the ACM 8:627-633 doi:10.1145/365628.365657

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

Sánchez D, Batet M (2012) A new model to compute the information content of concepts from taxonomic knowledge International Journal on Semantic Web and Information Systems 8:34-50

Sánchez D, Batet M (2017) Toward sensitive document release with privacy guarantees Engineering Applications of Artificial Intelligence 59:23-34

Sánchez D, Batet M, Isern D (2011) Ontology-based information content computation Knowledge-based Systems 24:297-303

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

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

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

Sebti A, Barfroush AA (2008) A new word sense similarity measure in WordNet. Paper presented at the Proc. of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008, Wisia, Poland,

Seco N, Veale T, Hayes J An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In: López de Mántaras R, Saitta L (eds) 16th European Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27 2004. IOS Press, pp 1089-1090

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

Viejo A, Sánchez D (2016) Enforcing transparent access to private content in social networks by means of automatic sanitization Expert Systems with Applications 62:148-160

Viejo A, Sánchez D, Castellà-Roca J (2012) Preventing automatic user profiling in Web 2.0 applications Knowledge-Based Systems 36:191-205

Wang P, Domeniconi C Building semantic kernels for text classification using wikipedia. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA, 2008. ACM, pp 713-721

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

Yuan Q, Yu Z, Wang K A new model of information content for measuring the semantic similarity between concepts. In: Proceedings of the Second International Conference on Cloud Computing and Big Data, 2013. IEEE Computer Society, pp 141-146

Zhou Z, Wang Y, Gu J A New Model of Information Content for Semantic Similarity in WordNet. In: Yau SS, Lee C, Chung Y-C (eds) Second International Conference on Future Generation Communication and Networking Symposia, FGCNS 2008, Sanya, Hainan Island, China, December 13-15 2008. IEEE Computer Society, pp 85-89. doi:10.1109/FGCNS.2008.16