A Semantic Similarity Metric Combining Features and Intrinsic Information Content

Giuseppe Pirró
gpirro@deis.unical.it

DEIS, University of Calabria, Rende (CS), Italy

Abstract

In many research fields such as Psychology, Linguistics, Cognitive Science and Artificial Intelligence, computing semantic similarity between words is an important issue. In this paper a new semantic similarity metric, that exploits some notions of the feature based theory of similarity and translates it into the information theoretic domain, which leverages the notion of Information Content (IC), is presented. In particular, the proposed metric exploits the notion of intrinsic IC which quantifies IC values by scrutinizing how concepts are arranged in an ontological structure. In order to evaluate this metric, an on line experiment asking the community of researchers to rank a list of 65 word pairs has been conducted. The experiment's web setup allowed to collect 101 similarity ratings and to differentiate native and non-native English speakers. Such a large and diverse dataset enables to confidently evaluate similarity metrics by correlating them with human assessments. Experimental evaluations using WordNet indicate that the proposed metric, coupled with the notion of intrinsic IC, yields results above the state of the art. Moreover, the intrinsic IC formulation also improves the accuracy of other IC-based metrics. In order to investigate the generality of both the intrinsic IC formulation and proposed similarity metric a further evaluation using the MeSH biomedical ontology has been performed. Even in this case significant results were obtained. The proposed metric and several others have been implemented in the Java WordNet Similarity Library.

Key words: Semantic Similarity, Intrinsic Information Content, Similarity Experiment, Similarity on Mesh, Java WordNet Similarity Library

1 Introduction

Assessing semantic similarity between words is a central issue in many research areas such as Psychology, Linguistics, Cognitive Science, Biomedicine, and Artificial Intelligence. Semantic similarity can be exploited to improve

accuracy of current Information Retrieval techniques (e.g., [12,8]), to discover mapping between ontology entities [21], to validate or repair ontology mappings [16], to perform word-sense disambiguation [23]. Recently Li and colleagues in [14] proposed a methodology to compute similarity between short sentences through semantic similarity. Semantic similarity has also found its way in the context of Peer to Peer networks (e.g., [5]) where it can be exploited to perform semantic-based query routing. In particular, concepts of a shared taxonomy can be exploited both to define peer expertise and express semantic queries. Semantic similarity allows to compute neighborliness on a semantic basis, that is, by computing similarity among peer expertises. The neighbors to route a given message to can be chosen by computing the semantic similarity between concepts in a query and those reflecting neighbors' expertises. In [25] several applications of similarity in Artificial Intelligence are discussed . Also in the biomedical domain there exist some applications to compute semantic similarity between concepts of ontologies such as Gene (e.g. [19,2]) with the aim to assess, for instance, protein functional similarity. However, despite the numerous practical applications of semantic similarity, it is important pointing out its theoretical underpinning in Cognitive Science and Psychology where several investigations (e.g.,[28]) and theories (e.g., [17,32]) have been proposed.

As a matter of fact, semantic similarity is relevant in many research areas and therefore, designing accurate methods is mandatory for improving the "performance" of the bulk of applications relying on it. Basically, similarity or distance methods aim at assessing a score between a pair of words by exploiting some information sources. These can be search engines (e.g., [1,3]) or a well-defined semantic network such as WordNet [18] or MeSH [1]. To date, several approaches to assess similarity have been proposed, which can be classified on the basis on the source of information they exploit ([9] provide an exhaustive list of references). *Ontology-based* approaches (e.g., [22]) assess semantic similarity by counting the number of nodes/edges separating two concepts. Even if these strategies are the most intuitive and easy to implement they suffer from the limitation that to work properly require consistent and rich ontologies, that is, ontologies where the leap between general concepts and that between specific ones have the same interpretation. *Information-theoretic* approaches (e.g., [15,10,24]) exploit the notion of Information Content (IC) defined as a measure of the informativeness of concepts and computed by counting the occurrence of words in large corpora. The drawbacks here is that it is necessary to perform time-consuming analysis of corpora and that IC values can depend on the kind of the considered corpora. *Hybrid* approaches (e.g., [13,34]) combine multiple information sources. A limitation of these approaches is that typically require some "configuration knobs" (e.g., weights used to set the contribution of each information source) to be adjusted.

[1] `http://www.nlm.nih.gov/mesh`

The purpose of this paper is to systematically design, evaluate and implement a new similarity metric to solve the shortcomings of existing approaches. In particular, the new similarity metric (named as P&S) exploits some of the early work done on the feature-based theory of semantic similarity proposed by Tversky [32], and projects it into the information theoretic domain. The P&S metric has not been derived empirically but has a theoretical underpinning in the feature-based theory of semantic similarity. As the extensive experimental evaluation performed will show (see sections 5 and 6), this metric coupled with the notion of *intrinsic* Information Content [30] outperforms current implementations on different datasets. Besides, the P&S metric neither require complex IC computations nor configuration knobs to be adjusted.

In order to evaluate the proposed and other metrics, a similarity experiment to collect ratings of similarity provided by human has been conducted. The number of participants in the present experiment is significantly higher than that of previous experiments and hence it hopefully will provide a more robust and reliable evaluation tool. Moreover, by correlating the collected ratings with those collected by the previous R&G experiment [27], an interesting investigation on the possible upper-bound for results that we can expect from computational methods has been conducted. In order to evaluate the generality of both the intrinsic IC formulation and proposed metric, a twofold evaluation has been performed. In both cases the P&S metric obtained results above the state of the art. To give more credit to the evaluations statistical significance tests have also been performed. Finally, the P&S metric and several others have been implemented in the Java WordNet Similarity Library [2], which is one of the few tools written in Java devoted to compute similarity in WordNet.

The remainder of this paper is organized as follows. Section 2 provides some background information regarding WordNet and popular similarity metrics. Here pros and cons of the state of the art will be highlighted with the aim to motivate the metric devised in this paper. Section 3 presents the P&S similarity metric and the intuitions that motivated its origin. In Section 4 how the new dataset used in the evaluation was created and its comparison w.r.t previously used datasets are discussed. Section 5 uses the new dataset to analyze and compare several similarity metrics, by correlating them to the human assessments. Moreover, here the impact of the *intrinsic* IC formulation on similarity metrics is discussed. This section also discusses a new upper bound on the degree of correlation that may be obtained using computational approaches. In Section 6 the generality of both the intrinsic IC formulation and similarity metric are investigated. In Section 7 possible extensions of the proposed metric are discussed. Finally, Section 8 concludes the paper.

[2] `http:\\grid.deis.unical.it\similarity`

2 WordNet and Similarity Metrics

WordNet is a light-weight lexical ontology where concepts are connected to each other by well-defined types of relations. It is intended to model the human lexicon, and took psycholinguistic findings into account during its design [17]. We call it a light-weight ontology because, despite having several types of lexical relations, it is heavily grounded on its taxonomic structure that employs the IS-A inheritance relation. Fig. 1 shows an excerpt of the WordNet noun taxonomy. In WordNet concepts are referred to by different words; for example if we want to refer to the concept expressed by *"someone deranged and possibly dangerous*' we could use any of the words contained in the set $\{Crazy, Loony, Looney, Weirdo\}$. So in a given context we can say that the words in the above set are synonyms. Hence, a synset (Synonym Set), the term adopted by the founders of WordNet, represents the underlying lexical concept. Each concept contains a gloss that expresses its semantics by means of a textual description and a list of words that can be used to refer to it. There are several types of relations used to connect the different synsets. Some of these define inheritance (*IS-A*) relations (i.e., *hypernymy/hyponymy*), other *part-of* relations (i.e., *holonymy/meronymy*). The *antonymy* relation is used to state that a noun is the opposite of another. The relations *instance of* and *has instance* are used to define instances. However, it is worth noting that the *hypernymy/hyponymy* relations constitute the majority of the relations connecting noun synsets.

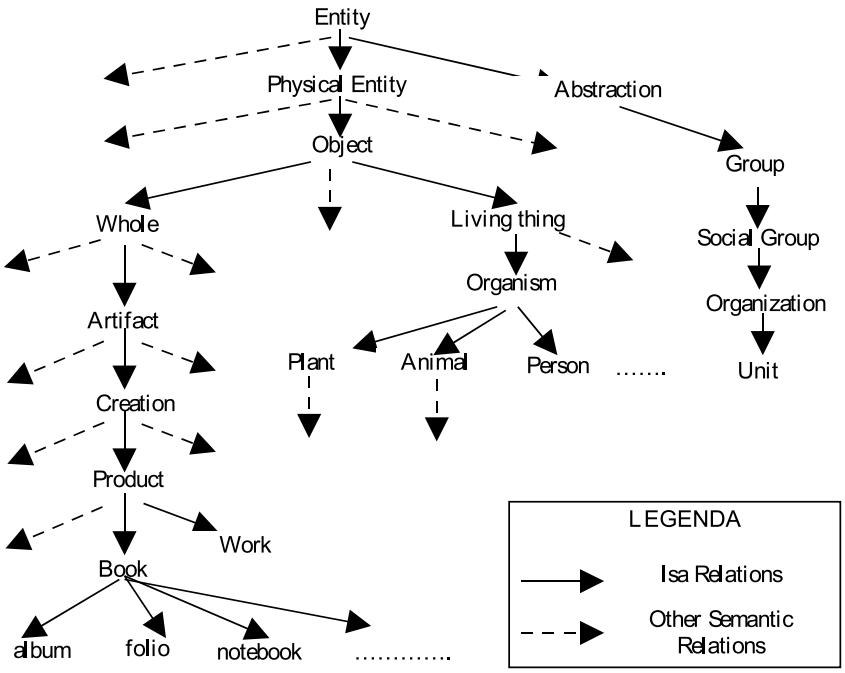


Fig. 1. An excerpt of the WordNet noun taxonomy

The prototypical definition of a noun consists of its immediate superordinate followed by a relative clause that describes how this instance differs from all other instances. For example, *Fortified Wine* is distinguished from *Wine* because *"... alcohol (usually grape brandy)"* has been added just as the gloss mentions. This type of model is usually said to employ a differential theory of meaning, where each subordinate differentiates itself from its super ordinate.

2.1 Similarity Metrics on WordNet

Similarity metrics between concepts can be divided into various, and not necessarily disjoint, categories [33]. In this paper we will focus on popular metrics that belong to the information theoretic, ontology-based or hybrid category. A complete survey of existing metrics is out of the scope of this paper (refer to [9] for a comprehensive list of references). In the following we review the state of the art metrics belonging to the abovementioned categories.

2.2 Information theoretic approaches

Information theoretic approaches usually employ the notion of Information Content (IC), which can be considered a measure quantifying the amount of information a concept expresses. Previous information theoretic approaches [24,10,15] obtained the needed IC values by associating probabilities to each concept in the taxonomy based on word occurrences in a given corpus. These probabilities are cumulative as we go up the taxonomy from specific concepts to more abstract ones. This means that every occurrence of a noun in the corpus is also counted as an occurrence of each taxonomic class containing it. The IC value is obtained by considering negative the log likelihood:

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

where c is a concept in the considered ontology and $p(c)$ is the probability of encountering c in a given corpus. It should be noted that this method ensures that IC is monotonically decreasing as we move from the leaves of the taxonomy to its roots. In fact, the concept corresponding to the root node of the IS-A hierarchy has the maximum frequency count, since it includes the frequency counts of every other concept in the hierarchy. Resnik [24] was the first to consider the use of this formula, which stems from the work of Shannon [31], for the purpose of semantic similarity judgments. The basic intuition behind the use of the negative likelihood is that the more probable a concept is of appearing the less information it conveys, in other words, specific words are more informative than general ones. Knowing the IC values for each concept we may then calculate the similarity between two given concepts.

According to Resnik, the similarity depends on the amount of information two concepts have in common. This shared information is given by the Most Specific Common Abstraction (*msca*) that subsumes both concepts. As an example, in Fig. 1 the concept *Organism* subsumes both *Plant* and *Person*. In order to find a quantitative value of the shared information we must first discover the *msca*. If one does not exist then the two concepts are maximally dissimilar, otherwise the shared information is equal to the IC value of their *msca*. Resnik's formula is modeled as follows:

$$sim_{res}(c_1, c_2) = max_{c \in S(c_1,c_2)} IC(c) \ . \tag{2}$$

where $S(c_1,\ c_2)$ is the set of concepts that subsume c_1 and c_2. Following Resnik's first work two other distinguishable metrics were postulated, that of Lin [15] and the work of Jiang and Conrath [10]. Both metrics used the notion of IC and calculated it in the same manner proposed by Resnik. Both Lin's and Jiang's formulations correct some problems with Resnik's similarity metric. First, if one were to calculate $sim_{res(c_1,c_1)}$ one would not obtain the maximal similarity value of 1, but instead the value given by IC(c_1). Second, with Resnik's metric any two pairs of concepts having the same *msca* have exactly the same semantic similarity. For example, $sim_{res}(Person, Plant) = sim_{res}(Animal, Plant)$ because in each case the *msca* is *Living Thing* (see Fig. 1).

According to Lin *"The similarity between c_1 and c_2 is measured by the ratio between the amount of information needed to state the commonality of c_1 and c_2 and the information needed to fully describe what c_1 and c_2 are"*. Formally this formula is given in the following equation:

$$sim_{Lin}(c_1, c_2) = \frac{2 \cdot sim_{res}(c_1, c_2)}{IC(c_1) + IC(c_2)} \ . \tag{3}$$

The Jiang et al. metric is a semantic distance measure and is derived from the edge-based notion of distance with the addition of the IC as a decision factor. As shown in [29] this distance metric can be transformed to a similarity metric yielding:

$$sim_{J\&C}(c_1, c_2) = 1 - \frac{IC(c_1) + IC(c_2) - 2 \cdot sim_{res}(c_1, c_2)}{2} \ . \tag{4}$$

2.3 Ontology based approaches

Regarding the ontology based approaches we review two noteworthy initiatives, one of Rada et al. [22] and the other of Hirst et al. [6]. The first is also referred to as a depth based approach and the second as a path based

approach. The Rada metric is similar to the Resnik metric in that it also computes the *msca* between two concepts, but instead of considering the IC as the value of similarity, it considers the number of links that were needed to attain the *msca*. Obviously, the less number of links separating the concepts the more similar they are. The approach of Hirst et al. [3] is similar to the previous but instead they use all types of relations in WordNet coupled with rules that restrict the way concepts are transversed. Nonetheless, the intuition is the same; the number of links separating two concepts is inversely proportional to the degree of similarity.

2.4 Hybrid approaches

Hybrid approaches combine different sources of information to assess a score of similarity or distance between concepts.

An approach combining structural semantic information in a nonlinear model is that proposed by Li et al. [13]. The authors empirically defined a similarity measure that uses shortest path length, depth and local density in a taxonomy. The next equation reflects their metric:

$$sim_{Li}(c_1, c_2) = \begin{cases} e^{-\alpha l \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}} & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases} \tag{5}$$

where l is the length of the shortest path between c_1 and c_2 in the graph spanned by the IS-A relation, h is the level in the tree of the *msca* from c_1 and c_2. The parameters α and β represent the contribution of the shortest path length l and depth h. The optimal values for these parameters, determined experimentally, are: $\alpha = 0.2$ and $\beta = 0.6$ as discussed in [13].

In [34] the *OSS* semantic distance function combining *a-priori* scores of concepts with concept distance is proposed. *OSS* performs the following three steps to assess similarity between two concepts c_1 and c_2: (i) computing the score of the concepts; (ii) computing how much score has been transferred between the two concepts; (iii) transforming the transfer of score into a distance measure $D(c_1, c_2)$. In more detail, the *a-priori* score for a concept, which expresses how much a concept is preferred or friendly into a particular context, is computed by analyzing the ontology structure. It aims at assessing and capturing the knowledge intrinsically included in a concept definition by the ontology designer. Scores decreases as we travel up the ontology structure as well as difference between scores. The *OSS* similarity metric, obtained by

[3] This approach actually measures relatedness, but since similarity is a special case of relatedness (see [29])we consider it in our study

subtracting 1 to the distance metric, is shown in the next equation:

$$sim_{OSS}(c_1, c_2) = 1 - \frac{log(T(c_1, c_2))}{maxD} \ . \tag{6}$$

where $T(c_1, c_2)$ is the transfer of score from concept c_1 to c_2 and $maxD$ is the maximum distance between any two concepts in the ontology.

2.5 Comparison among metrics

Table 1 summarizes the peculiarities of the similarity metrics described in the previous sections. As can be noted, each metric has pros and cons. IC-based metrics making use of corpora, though having a strong mathematical formalization, may sometimes fail to capture certain aspects of language. The values of IC are obtained through time intensive analysis of corpora and can heavily depend from the considered corpora. Thus, it is possible that some large corpora, such as the British National Corpus, may not even mention certain words. As for ontology-based approaches, even if they are simple, it is mandatory to work with consistent ontologies, that is, ontologies where distance between specific and more general concepts have the same interpretation. As an example it is evident that the semantic leap between *Entity* and *Psychological Feature* is higher than that between *Canine* and *Dog* even in both couples are separated by one edge. Finally, hybrid approaches require the different information sources to be correctly "weighted".

This analysis shows that different strategies have been proposed during the years that, as will be shown in Section 5, have been approaching more and more human assessments of similarity. This analysis has been helpful in designing the P&S similarity metric. In particular, our study rests on the following principles:

- The P&S metric has to be supported by a theoretical underpinning and has not to be empirically derived.
- The P&S metric has to exploit the benefits from IC-based techniques which, to date, achieve the higher correlation w.r.t human judgments. However, it is mandatory to avoid their drawbacks, that is, the time intensive analysis of corpora and the dependence from the considered corpora.
- The suitability of the metric has to be assessed by comparing it w.r.t human judgments of similarity. In particular, we argue that it would be useful to collect a large number of judgments of similarity and use a large dataset.

In the next sections we elaborate on these aspects and show the path we followed to fulfill each of them.

Table 1
Overall view of the different similarity metrics

Category	*Measure*	*Basic principle*	*Pros*	*Cons*
IC-based	Resnik [24]	IC as a measure of similarity	Word occurrences in corpora as an empiric measure of informativeness	Only considers the msca
	Lin [15]	Extension of Resnik's metric	Universal measure	Requires analysis of corpora to compute IC
	J&C[10]	Extension of Resnik's metric	Combines distance with IC	Requires analysis of corpora to compute IC
Ontology based	Rada[22]	Count of edges between concepts	Simplicity	Requires consistent ontologies
	Hirst[6]	Measures relatedness of all parts of speech	Computes relatedness	WordNet specific
Hybrid	Li[13]	Combination of multiple information sources	Non-linear combination of information sources	Requires parameters to be settled
	OSS[34]	Combination of *a-priori* scores and distance	No external sources	Tuning is required

3 The *P&S* Similarity Metric

In this section we introduce our new similarity metric which is conceptually similar to the previous ones, but is founded on the feature-based theory of similarity posed by Tversky [32]. We argue that his theory fits nicely into the information theoretic domain, and obtains results that improve the current state of the art. The argumentation presented here follows from the work conducted in [20,29].

Tversky presented an abstract model of similarity, based of set theory, that

takes into account the features that are common to two concepts and also the differentiating features specific to each. As an example, since *car* and *bicycle* both serve to transport people or objects, in other words they are both types of vehicles, they share all features that pertain to the concept *vehicle*. However each concept has also its specific features as *steering wheel* for *car* and *pedal* for *bicycle*.

According to Tversky, the similarity of a concept c_1 to a concept c_2 is a function of the features common to c_1 and c_2, those in c_1 but not in c_2 and those in c_2 but not in c_1. Figure 2 provides a graphical representation of Tversky's model.

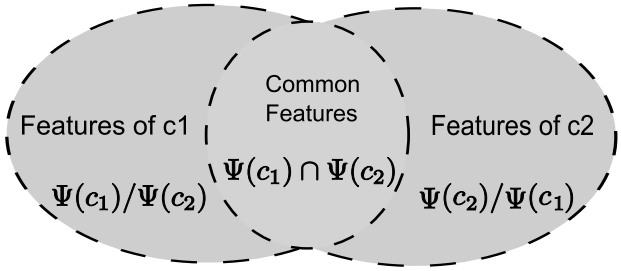


Fig. 2. Tversky's similarity model

Admitting a function $\psi(c)$ that yields the set of features relevant to c, he proposed the following similarity function:

$$sim_{tvr}(c_1, c_2) = \alpha \cdot F(\Psi(c_1) \cap \Psi(c_2)) - \beta \cdot F(\Psi(c_1)/\Psi(c_2)) - \gamma \cdot F(\Psi(c_2)/\Psi(c_1)) \ . \tag{7}$$

where F is some function that reflects the salience of a set of features, and α, β and γ are parameters that provide for differences in focus on the different components. According to Tversky, similarity is not symmetric, that is, $sim_{tvr}(c_1, c_2) \neq sim_{tvr}(c_2, c_1)$ because subjects tend to focus more on one object than on the other depending on the way the comparison experiment has been laid out.

Obviously, the above formulation is not framed in information theoretic terms. Nonetheless, we argue that a parallel may be established that will lead to a new similarity function. Resnik considered the *msca* of two concepts c_1 and c_2 as reflecting the information these concepts share, which is exactly what is intended with the intersection of features from c_1 and c_2 (i.e., $\Psi(c_1) \cap \Psi(c_2)$). Now, remembering that function F quantifies the salience of a set of features, then we postulate that we may find that quantification in the form of information content. The above reasoning will lead us to the analogy

represented in the following equation:

$$\begin{aligned} sim_{res}(c_1, c_2) =& IC(msca(c_1, c_2)) \approx F(\Psi(c_1) \cap \Psi(c_2)) \\ =& F(\Psi(c_1) \cap \Psi(c_2)) - 0 \cdot F(\Psi(c_1)/\Psi(c_2)) - 0 \cdot F(\Psi(c_2)/\Psi(c_1)) \ . \end{aligned} \tag{8}$$

Since the *msca* is the only parameter taken into account we may say that his formulation is a special case of equation (7) where $\beta = \gamma = 0$. The above discussion lends itself to the proposal of an information theoretic counterpart of equation (7) that can be formalized as:

$$\begin{aligned} sim_{tvr'}(c_1, c_2) =& IC(msca(c_1, c_2)) - (IC(c_1) - IC(msca(c_1, c_2)) + \\ & - (IC(c_2) - IC(msca(c_1, c_2)) \\ =& 3 \cdot IC(msca(c_1, c_2)) - IC(c_1) - IC(c_2) \ . \end{aligned} \tag{9}$$

Note that the equality $\Psi(c_1)/\Psi(c_2) = IC(c_1) - IC(msca(c_1, c_2))$ indicates that the IC of the concept c_1 is obtained by subtracting from its features the common features $\Psi(c_1) \cap \Psi(c_2)$ as shown in Fig. 2. Same considerations are valid for the equality $\Psi(c_2)/\Psi(c_1) = IC(c_2) - IC(msca(c_1, c_2))$.

A careful analysis of equation (9) shows that this metric suffers from the same problem as Resnik's metric; when computing the similarity between identical concepts the output yields the information content value of their *msca* and not the value corresponding to maximum similarity (i.e., the value 1 obtained when comparing a concept with itself). In order to overcome this limitation we assign the value of 1 if the two concepts are the same, hence yielding the similarity metric that can be formalized as follows:

$$sim_{P\&S}(c_1, c_2) = \begin{cases} sim_{tvr'} & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases} \tag{10}$$

Note that in equation (10) we use $sim_{tvr'}$ which is the information theoretic counterpart of Tversky's set theoretic formulation. This new formulation will dubbed as the $P\&S$ metric in the rest of the paper.

Remark. The formulation of the $P\&S$ metric given in equation (10) exploits the strengths of IC-based approaches, corrects the problem of Resnik metric (i.e., $sim_{Res}(c_1, c_1) \neq 1$) while at the same time having the feature-based theoretical underpinning. Moreover, this metric does not require parameters to be adjusted. At this point only a possible drawback related to IC-metrics remains to be solved: how to obtain IC values in a more direct and corpus-independent way ? This problem is addressed in the next section.

As pointed out before, the conventional information theoretic way of measuring information content of word senses is to combine knowledge of their hierarchical structure from an ontology such as WordNet with statistics on their actual usage in text as derived from a large corpus (e.g., [24]). The intuitive motivation behind this type of reasoning is that rare concepts are more specific and therefore much more expressive. However, from a practical point of view, this approach has two main drawbacks:

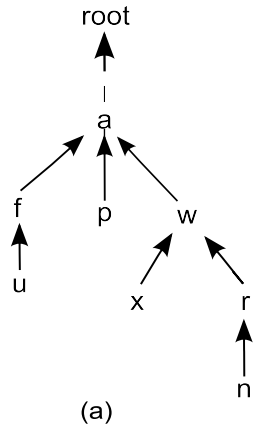
(1) It is time consuming since it implies that large corpora should be parsed and analyzed and the considered corpora have to be adequate in term of content for the considered ontology.
(2) It heavily depends on the type of corpora considered and its size. It is arguable that IC values obtained from very general corpora maybe different than those obtained with more specialized ones. As for WordNet, the Brown corpus is typically exploited, which is a good source of general knowledge. However, if we were calculating similarity between concepts of very specialized ontologies such as MeSH, it is likely that Brown corpus does not contain many of the terms included in that ontology and then IC values and corresponding similarity assessments could be affected.

Research toward mitigating these drawbacks has been proposed by Seco et al. [30]. Here, values of IC of concepts rest on the assumption that the taxonomic structure of the ontology (e.g.,WordNet) is organized in a meaningful and structured way, where concepts with many hyponyms convey less information than concepts that are leaves. The intuition is: more abstract concepts are more probable of being present in a corpus because they subsume so many other ones. Given this, it is possible to speculate on the probability of a concept to appear by considering the number of hyponyms it has. If a concept has many hyponyms, then it has more of a chance of appearing since the subsuming concept is implicitly present when reference to one of its hyponyms is made. So if one wanted to calculate the probability of a concept it would be the number of hyponyms it has plus one (for itself) divided by the total number of concepts that exist. The *intrinsic* IC for a concept c is defined as:

$$IC(c) = 1 - \frac{log(hypo(c) + 1)}{log(max_{con})} \; . \tag{11}$$

where the function *hypo* returns the number of hyponyms of a given concept c. Note that concepts representing leaves in the taxonomy will have an IC of one, since they do not have hyponyms. The value of one states that a concept is maximally expressed and can not be further differentiated. Moreover max_{con} is a constant that indicates the total number of concepts in the considered taxonomy.

The intrinsic IC formulation is based on the assumption that the ontology is organized according to the principle of *cognitive saliency* [33]. Cognitive saliency states that humans create concepts when there is a need to differentiate from what already exists. As an example, in WordNet the concept *cable car* only exists because its lexicographers agree that *cable* is a sufficiently salient feature [4] allowed it to be differentiated from *car* and promoted to a concept in its own right. Obviously, what is cognitively salient to one community may not be to another and consequently these communities will have different similarity judgments. WordNet seeks to be a general purpose lexical ontology, trying to cover lexical concepts from as many domains as possible and then it can be argued that WordNet is constructed following principles of general knowledge [29]. We will evaluate the impact of the intrinsic IC formulation on WordNet and show in Section 6 that this formulation achieves acceptable results even considering more specialized ontologies (e.g., MeSH). Fig. 3 shows an example of intrinsic IC calculation. Fig. 3 (a) shows an ontology structure while intrinsic IC values are shown in Fig.3 (b). As can be noted, IC values decreased as we travel up the taxonomy (e.g., $IC(a) < IC(r)$).



Concept	Intrinsic IC
root	0
a	0.04
f	0.1
p	1
w	0.052
u	1
x	1
r	0.1
n	1

(b)

Fig. 3. An example of IC calculation

This definition of IC will be exploited in the $P\&S$ similarity metric thus enabling to solve the last problem previously highlighted. In Section 5.5 we show how intrinsic IC improves the accuracy of other IC based metrics as well.

4 The $P\&S$ Similarity Experiment

In order to assess the quality of a computational method to determine similarity between words, that is, its accuracy, a natural way is to compare its

[4] This feature indicates that the *cable car* operates on a cableway or cable railway as the WordNet gloss mentions.

behavior w.r.t human judgments. The more a method approaches human similarity judgment the more accurate it is.

In evaluating the different methodologies two datasets are commonly used, those of Rubinstein and Goodenough (R&G in the following) and Miller and Charles (M&C in the following). R&G [27] in 1965 performed a similarity experiment by providing 51 human subjects, all native English speakers, with 65 word pairs and asking them to assess similarity between word pairs on a scale from 0 ("semantically unrelated") to 4 ("highly synonymous"). M&C [18], 25 years later, repeated the R&G experiment by only considering a subset of 30 words pairs from the original 65, and involving 38 undergraduate students (all native English speakers). In this case humans were also asked to rate similarity between pairs of words on a scale from 0 to 4. Although the M&C experiment was carried out 25 years later, the correlation between the two sets of human ratings is 0.97 which is a very remarkable value considering the diachronic nature of languages. Resnik [24] on his turn in 1995 replicated the M&C experiment by involving 10 computer science graduate students and post-doc (all native English speakers) obtaining a correlation of 0.96, also in this case a high value.

The results of these experiments point out that human knowledge about semantic similarity between words is remarkably stable over years (25 and 30 years later the R&G, for the M&C and Resnik experiment respectively). Moreover, they also point out how the usage of human ratings could be a reliable reference to compare computational methods with. However, researches tend to focus on the results of the M&C experiment to evaluate similarity metrics and, to the best of our knowledge, no systematic replicas of the entire R&G experiment have been performed. Therefore, we argue that it would be valuable to perform a "new" similarity experiment in order to obtain a baseline for comparison with the entire R&G dataset.

4.1 Experiment Setup

We replicate the R&G experiment (naming it P&S in the following) but one step closer to the 21st century, the century of the Internet and global information exchange. In particular, we performed the experiment on the Internet by advertising it in some of the most famous computer science mailing lists (e.g., DBWORLD, CORPORA, LINGUIST) with the aim to involve as many people as possible. Each participant, after a registration process on the similarity experiment website [5] could take part in the experiment. In the web site were provided all the instructions to correctly perform the experiment. The similarity scores along with the emails provided by participants have been stored

[5] The similarity experiment web site: `http://grid.deis.unical.it/similarity`

in a database for subsequent analysis. As one can imagine, and as our results confirmed, the participants were mostly graduate students, researchers and professors. Note that we also opened the experiment to non native English speakers. As said above, in the era of global information exchange more and more people speak English thus participating in the creation and spreading of new forms of interpreting terms. Furthermore, semantic relations among words are affected by language evolution that, on its turn, is affected by the presence of a larger number of speakers of a particular language. Our objective is to investigate if and how the presence of non native speakers affects similarity judgments.

In particular, in our experiment about 70% of native speakers are American English speakers, 30% are British English speakers while non native speakers are for the most part European. Table 2 provides some information about the experiment. As can be noted, even if we collected 121 similarity ratings we discarded some of them for the reasons explained in the next section.

Table 2
Information about the $P\&S$ experiment

Start of the experiment	07/15/2007
Result considered until	04/15/2008
# of similarity judgments collected	121
# of similarity judgments considered in the gold standard	101
# of similarity judgments provided by native speakers	76
# of similarity judgments provided by **non** native speakers	25

4.2 Elaborating the Collected Similarity Ratings

In order to design a systematic experiment and consider its results reliable, an a posteriori analysis of its results is required. In our case, this analysis is particularly important for similarity ratings provided by non native speakers since the group of non native speakers could be quite large and heterogeneous, ranging from near-native speakers over very fluent speakers to speakers with only rudimentary knowledge of English. In order to check the quality of the ratings provided by the participants, we calculated, for each participant, a rating coefficient (i.e., C) defined as follows:

$$C = \sum_{i=1}^{65} |C_i - avg_i| \ . \tag{12}$$

In particular, for each word pair the distance between the score provided by the participant and the average score provided by the others is measured. The distance values for all the 65 pairs are then summed up. Once computing all the coefficients C we could discard the participants that present values of C differing too much from the average. Fig. 4 represents the C values for all the 121 participants.

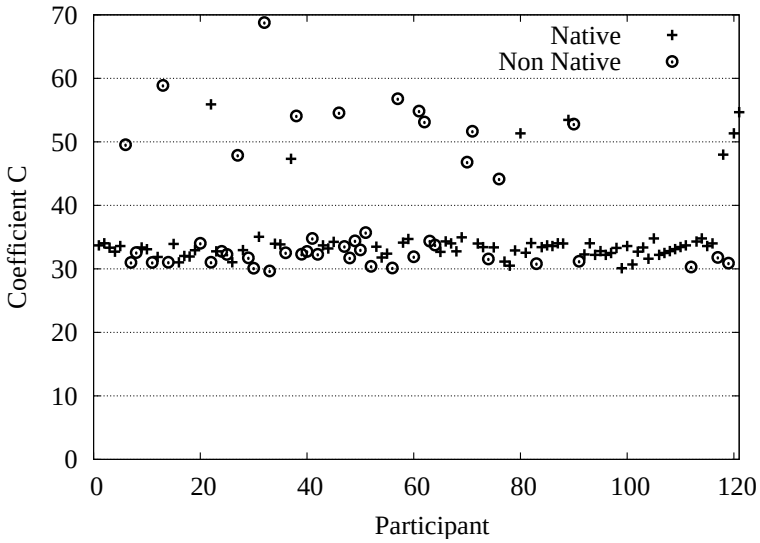


Fig. 4. Values of the coefficient C for the participants to the P&S experiment

As can be noted, most of the C coefficients lie between 30 and 40. However, ratings provided by some participants (and then C coefficients) clearly differers from the average. The ratings provided by these participants have been discarded. In particular, by observing the results provided in Fig. 4 it can be noted that the anomalous ratings were for the most part given by non native speakers (about 65%). Table 3 provides an overall view of the different similarity experiments.

Table 3
Overall view of the different similarity experiments

Experiment	Year	Number of pairs	Number of participants
R&G	1965	65	51 (all native speakers)
M&C	1991	30	38 (all native speakers)
Resnik	1995	30	10 (all native speakers)
P&S	2008	65	101(76 native and 25 non native)

Note that even if we collected 121 similarity ratings, we only considered 101 as reliable. We collected a larger number of similarity ratings than R&G, M&C and Resnik experiments and about 30% of participants in our experiment are (reliable) non native English speakers. Moreover, differently from M&C and

Resnik we performed the experiment by considering the whole initial R&G dataset.

In Table 4 the average ratings of similarity provided by the 101 human subjects for the 65 R&G word pairs are reported. These results are compared with those obtained by the R&G experiment that involved 51 human subjects. Each experiment is represented in a column, in particular, P&S considering non native speakers and P&S considering only native speakers are represented in columns $P\&S_{full}$ and $P\&S_{nat}$ respectively. The 28 pairs in the $S_{M\&C}$ are reported in bold.

Table 4
Similarity ratings for the P&S and R&G experiments considering the 65 word pairs

Pair	R&G	P&S$_{\text{full}}$	P&S$_{\text{nat}}$	Pair	R&G	P&S$_{\text{full}}$	P&S$_{\text{nat}}$
gem-jewel	3.940	3.563	3.296	*crane-rooster*	1.410	1.271	1.253
midday-noon	3.940	3.247	3.270	*furnace-implement*	1.370	1.111	1.168
automobile-car	3.920	3.421	3.544	**coast-hill**	1.260	0.961	0.966
cemetery-graveyard	3.880	3.430	3.490	*bird-woodland*	1.240	0.878	0.880
cushion-pillow	3.840	3.134	2.995	*shore-voyage*	1.220	0.931	0.960
boy-lad	3.820	3.026	3.103	*cemetery-woodland*	1.180	0.697	0.677
cock-rooster	3.680	3.143	3.203	**food-rooster**	1.090	1.048	1.061
implement-tool	3.660	2.644	2.654	**forest-graveyard**	1.000	0.801	0.789
forest-woodland	3.650	2.975	2.996	**lad-wizard**	0.990	0.752	0.710
coast-shore	3.600	3.026	3.016	*mound-shore*	0.970	0.753	0.743
autograph-signature	3.590	2.676	2.734	*automobile-cushion*	0.970	0.584	0.564
journey-voyage	3.580	2.945	3.028	*boy-sage*	0.960	0.585	0.539
serf-slave	3.460	2.554	2.551	**monk-oracle**	0.910	1.166	1.152
grin-smile	3.460	2.713	2.715	*shore-woodland*	0.900	0.756	0.742
glass-tumbler	3.450	2.504	2.483	*grin-lad*	0.880	0.589	0.628
cord-string	3.410	2.614	2.733	**coast-forest**	0.850	0.733	0.725
hill-mound	3.290	2.535	2.501	*asylum-cemetery*	0.790	0.672	0.701
magician-wizard	3.210	2.885	2.857	**monk-slave**	0.570	0.695	0.685
furnace-stove	3.110	2.325	2.386	*cushion-jewel*	0.450	0.582	0.583
asylum-madhouse	3.040	2.658	2.699	*boy-rooster*	0.440	0.643	0.648
brother-monk	2.740	1.997	1.997	**glass-magician**	0.440	0.509	0.611
food-fruit	2.690	1.757	1.794	*graveyard-madhouse*	0.420	0.562	0.550
bird-cock	2.630	1.661	1.714	*asylum-monk*	0.390	0.721	0.758
bird-crane	2.630	1.684	1.722	*asylum-fruit*	0.190	0.393	0.439
oracle-sage	2.610	2.216	2.222	*grin-implement*	0.180	0.515	0.561
sage-wizard	2.460	1.893	1.893	*mound-stove*	0.140	0.494	0.493
brother-lad	2.410	1.525	1.550	*automobile-wizard*	0.110	0.461	0.518
crane-implement	2.370	1.356	1.383	*autograph-shore*	0.060	0.450	0.537
magician-oracle	1.820	1.590	1.652	*fruit-furnace*	0.050	0.438	0.502
glass-jewel	1.780	1.146	1.156	**noon-string**	0.040	0.443	0.483
cemetery-mound	1.690	1.253	1.257	**rooster-voyage**	0.040	0.421	0.435
car-journey	1.550	1.176	1.278	**cord-smile**	0.020	0.476	0.482
hill-woodland	1.480	1.066	1.103				

In the collected ratings we noted that the couples of words *rooster-voyage*,*cord-smile* and *brother-monk* present the three highest standard deviations. This may affect the reliability of the data when using it as a basis to correlate computational methods against. This consideration has led to an investigation, discussed in Section 5.4, on how these problematic pairs can affect the performance of the P&S metric.

4.3 Comparison among Experiments

We split the collected similarity judgments in two sets. The first set ($S_{M\&C}$ in the following) contains the judgments for the 28 word pairs in the M&C experiment. These pairs are indicated in bold in Table 4. The second set ($S_{R\&G}$ in the following) contains the 65 word pairs in the R&G dataset. In particular, this latter dataset is used to define a possible upper-bound for computational methods to assess semantic similarity. Note that the word pairs in M&C, extracted from the original R&G dataset, are chosen in a way that they range from "highly synonymous" (e.g., *car-automobile*) to "semantically unrelated" (i.e., *cord-smile*) according to common-sense.

In tables 5 and 6 the Pearson correlation coefficient [4] among the different experiments are reported. As can be noted, the correlation values obtained by our experiment are high. In particular, the correlation values considering only native ($P\&S_{nat}$) and all the participants ($P\&S_{full}$) are almost the same. Therefore, we argue that results of the $P\&S$ experiment can be adopted as a reliable basis for comparing similarity metrics against. Moreover, since the number of judgments collected is larger than that collected by previous experiments and the presence of non native speakers does not affect the similarity judgments we hope to provide a more reliable and robust evaluation tool.

Table 5
Correlation on $S_{M\&C}$

	$P\&S_{full}$	$P\&S_{nat}$
$R\&G$(1965)	0.961	0.964
$M\&C$(1991)	0.951	0.955
Resnik(1995)	0.970	0.972

Table 6
Correlation on $S_{R\&G}$

	$P\&S_{full}$	$P\&S_{nat}$
$R\&G$(1965)	0.972	0.971

4.4 Inter-annotator Agreement and Correlation between Groups of Participants

In order to substantiate data collected by the P&S experiment, it becomes mandatory to estimate the degree to which it can be unduly affected by the subjective judgment of the participants. Such estimation is provided by the coefficients of inter-rater agreement (aka kappa statistic). In our experiment, a further important parameter is the correlation between ratings provided by native and non native speakers. Table 7 reports these values for the $S_{R\&G}$and $S_{M\&C}$ datasets.

Considering the $S_{M\&C}$ the kappa-statistic obtained is 0.820 which symbolizes

Table 7
Kappa-statistic and correlation between groups of raters for $S_{R\&G}$ and $S_{M\&C}$

	Kappa-statistic	*Correlation between groups of raters*
$S_{M\&C}$	0.820	0.970
$S_{R\&G}$	0.810	0.980

the agreement among participants in rating the word pairs. On the same dataset, the correlation between the average judgments of native and non native speakers is 0.970, which is a very high value. Considering the $S_{R\&G}$ the kappa-statistic obtained is 0.810 while the correlation between the average judgments of non native and native speakers in this case is 0.980. Finally, note that the experiments involved a different number of participants (51 for R&G, 30 for M&C, 10 for Resnik and 101 for P&S).

5 Evaluation and Implementation of the P&S Metric

In this section, to substantiate the investigation that led to the definition of the P&S metric, we evaluate and compare it w.r.t the state of the art. In performing this evaluation we consider the results of the $P\&S$ experiment on the $S_{M\&C}$ and $S_{R\&G}$ datasets. All the evaluations have been performed using WordNet 3.0.

5.1 Evaluation methodology

In order to evaluate the accuracy of computational methods for assessing semantic similarity a commonly accepted approach is that of correlating their results with those obtained by humans performing the same task. We adopt the Pearson correlation coefficient as a measure of the *strength* of the relation between human ratings of similarity and computational values. However, to have a deeper interpretation of the results we also evaluate the *significance* of this relation. To this aim, we adopt the classical *p-value* approach, which tells how unlikely a given correlation coefficient, r, will occur given no relation in the population. Note that the smaller the *p-level*, the more *significant* the relation. Conversely, the larger the correlation the *stronger* the relation. The *p-value* for Pearson's correlation coefficient is based on the test statistic s defined as follows:

$$s = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}} \tag{13}$$

where r is the correlation coefficient and n is the number of pairs of data. When the null hypothesis (i.e., H_o) is true (i.e., when the population correlation

coefficient is equal to zero), the s statistic above follows a t distribution with n - *2* degrees of freedom [11]. P-values have been calculated through the Minitab statistical software [6].

5.2 Evaluation of the P&S metric

In our evaluation, we represent similarity values obtained by the different metrics as shown in Fig. 5. This way, we can discuss and characterize in detail the peculiarities, analogies and differences of the different metrics. However, to have an overall view of the outcome of our evaluation we calculated, for each metric, the Pearson correlation coefficient between its results and human judgments. Results are reported in tables 8 and 9.

Table 8
Correlation on $S_{M\&C}$

	P&S (2008)	
	$P\&S_{full}$	$P\&S_{nat}$
Length	0.611	0.602
Depth	0.841	0.839
Resnik	0.854	0.842
Lin	0.875	0.871
J&C	0.884	0.883
Li	0.911	0.904
P&S	0.912	0.908

Table 9
Correlation on $S_R\&G$

	P&S (2008)	
	$P\&S_{full}$	$P\&S_{nat}$
Length	0.587	0.578
Depth	0.807	0.805
Resnik	0.877	0.869
Lin	0.892	0.888
J&C	0.878	0.877
Li	0.900	0.897
P&S	0.908	0.905

The similarity values for the Length and Depth metrics are obtained by considering the shortest path between the two words to be compared and the depth of their subsumer respectively. For the metrics based on IC and the $P\&S$ metric the values of IC are obtained by the method described in Section 3.1. Moreover, for the Li metric the similarity results are those reported in [13].

The *p-values*, obtained by the method described in Section 5.1, in the two datasets are $p - value < 0.001$. This indicates that our results are highly significant.

[6] `http://www.minitab.com`

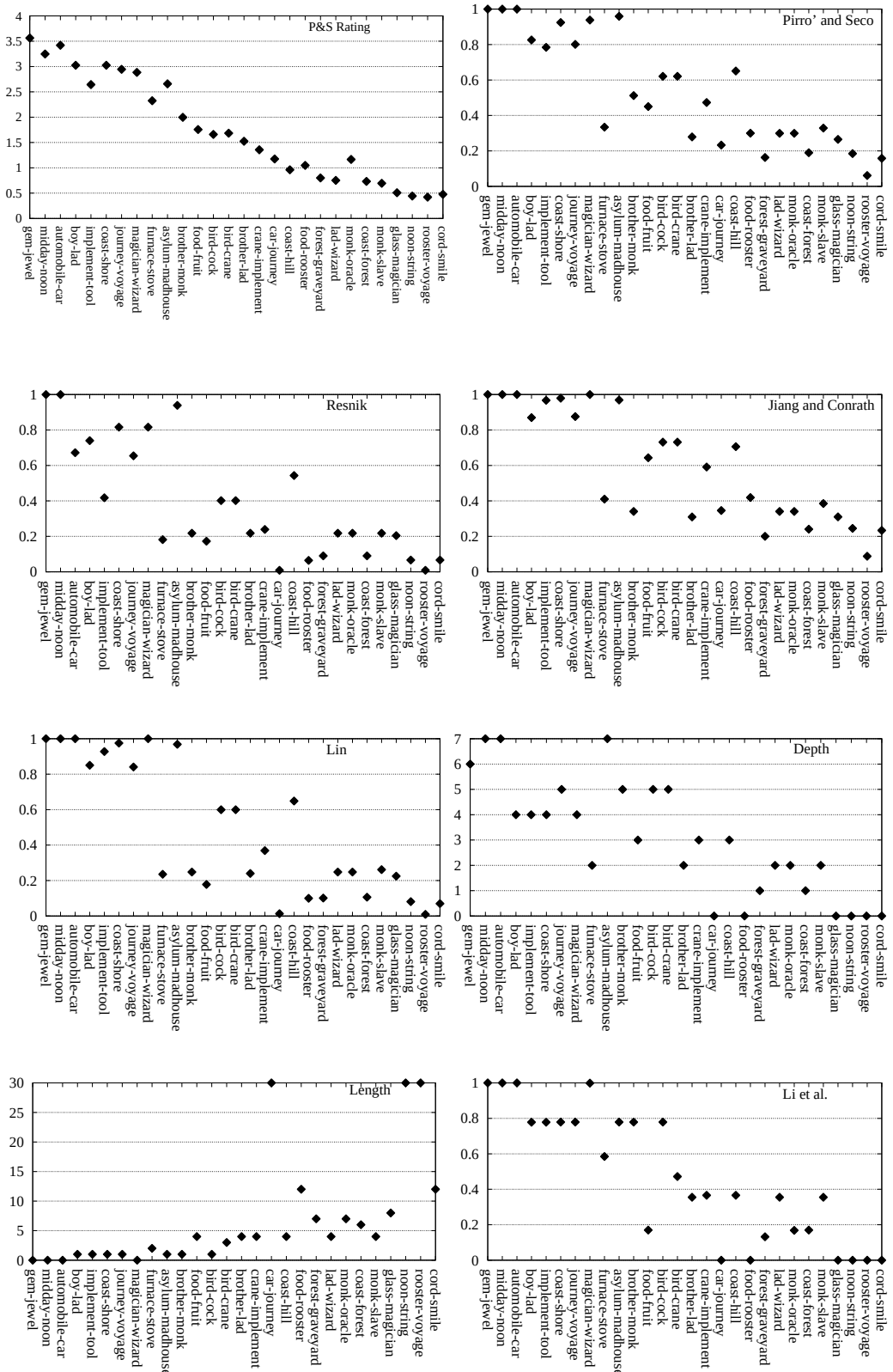


Fig. 5. Results and ratings considering the P&S dataset. Vertical axis represents the similarity value while horizontal axis word pairs

5.2.1 Discussion of the results

From the values reported in tables 8 and 9 it emerges that edge counting approaches reach the lowest correlation with human ratings. That is mainly due to the fact that path lengths and depth approaches are appropriate only when the values of path and depth have a "consistent interpretation". This is not the case of WordNet, since concepts higher in the hierarchy are more general than those lower in the hierarchy. Therefore, a path of length one between two general concepts can suggest a larger semantic leap whereas one between two specific concepts may not (e.g., *Entity − Psychological Feature* and *Canine − Dog*). Resnik's metric, which only considers the IC of the *msca* in assessing semantic similarity, obtained the lowest value of correlation among the IC metrics using $S_{M\&C}$. The Lin and $J\&C$ metrics, which consider also the IC of the two words to be compared, obtained higher values of correlation on $S_{M\&C}$. In particular, the Lin metric combines the information of the two concepts to be compared along with the information of their subsumer while the $J\&C$ metric combines an IC formulation of semantic similarity with edge counting. The Li metric obtained a remarkable value of correlation. Note that this metric which combines the depth of the *msca* and the length of the path between two concepts to be compared relies on two coefficients (i.e., α and β) whose optimal values have been experimentally determined as described in [13]. The $P\&S$ metric obtained a slightly higher value of correlation on the $S_{M\&C}$ dataset.

On the second dataset, that is $S_{R\&G}$, the correlation values obtained by the different metrics slightly change. Even in this case, the Length metric obtains the poorest correlation. Resnik's metric obtained a correlation comparable to that obtained by the $J\&C$ metric. The Lin metric obtained slightly better results. The Li metric, in this case evaluated by considering the optimal parameter determined by authors in [13] obtained a better correlation. However, the $P\&S$ metric remains the most correlated w.r.t human judgments also in this dataset. Correlation results reported in tables 8 and 9 show that the presence of non native speakers barely affects the values of correlation of the different metrics.

5.3 Commonalities and Differences among Metrics

In order to have a deeper insight into the structure of the different metrics, we represent their results as shown in Fig. 5. Here, it can be recognized the different nature of edge-counting (i.e., Length, Depth), IC-based and Li's multi-source metrics. In particular, edge- counting metrics give discrete results as output (i.e., integer values). For the Length metric, a low value of length corresponds to a higher similarity value between words. For instance the first

three pairs (i.e., *gem-jewel*, *midday-noon* and *automobile-car*) have a length equal to zero which is due to the fact that these word pairs belong to the same WordNet synset respectively. On the other side, word pairs as *noon-string* and *rooster-voyage* have a relatively high distance which means that the words in the two pairs are not similar. A potential anomaly could be represented by the pair *car-journey* which gets a length of 30, the maximum value. The two words, even if generally related as a car can be the means to do a journey, are not considered similar. That is because similarity is a special case of relatedness and only considers the relations of hypernymy/hyponymy defined in WordNet which is exactly what the Length metric does. For the Depth metric, a number of "similarity levels" can be recognized (in Fig. 5 for instance it can be noticed that there are 3 ratings in the level 7, 5 in the level 2 and 6 in the level 0). This metric, differently from that of Resnik takes into account the depth of the *msca* thus allowing more specific concepts to be generally judged more similar than more abstract one. Note that this metric obtained a correlation about 30% better than that obtained by the Length metric.

A more interesting discussion can be done for the IC based metrics. In particular, the Resnik and Lin metric present two similar regions, one in the center identified by the pairs *bird-cock* and *bird-crane* (translated by 0.2) and the other comprising all the pairs from *car-journey* to *cord-smile*. Note that when the two words to be compared are leaves, according to the intrinsic IC formulation described in equation (11), they have IC equals to 1 and therefore equation (3) turns into equation (2). A similar condition holds for the transformed J&C metric in equation (4). The $P\&S$ metric when c_1 and c_2 are leaves gives as result $sim_{P\&S}(c_1, c_2) = 3 \cdot IC(msca) - 2$. In this case, if the *msca* is high in the taxonomy (it receives a low IC) the metric returns a lower similarity value than when it is low. In particular, if the msca is very high (i.e., near the root concept) the P&S metric can give as output a negative value near -2 which can be interpreted as the maximum dissimilarity value.

A similar area can be recognized between the J&C and P&S metric (i.e., from the pairs *forest-graveyard* to *cord-smile*). In this area generally, the J&C obtains higher similarity scores. However, according to the original intent of R&G to chose word pairs from very similar to less similar, here the P&S metric seems to better respect this trend. Finally the Li metric has a very similar region (comprising the pair from *car-journey* to *cord-smile*) to the Depth metric. Word pairs in this region are rated equally due to the fact that the Li metric exploits the value of Depth and when this is zero, according to equation (5) the similarity value returned by the Li metric is zero.

In summary, the results of these experiments demonstrate that our intuition to consider the original formulation of IC provided by Resnik, to some extent, a special case of the formulation given by Tversky is consistent. Moreover, the metric (i.e., Li) that obtained results comparable to the P&S metric has been

empirically designed and relies on two parameters to be adjusted.

5.4 Problematic pairs

As discussed in Section 4.2, some of the collected similarity ratings present a value of standard deviation higher than that of the other pairs. In this section we investigate how these ratings may affect values of correlation for the P&S metric. In order to do that, each couple was assigned a score between a certain range, and then the corresponding value of correlation was computed. Table 10 shows the considered pairs along with the range of variation for their score. All experiments have been performed on the 65 word pairs in the $S_{R\&G}$. As for the evaluation methodology, we considered the pairs variating from one at time to all the three at the same time.

Table 10
P&S experiment problematic pairs

Pair	*Range of the score*	*Step*
rooster-voyage	0 - 0.5	0.01
cord-smile	0 - 0.5	0.01
brother-monk	1.8 - 2.8	0.05

The following figures report the value of correlation by considering the variation of score for each of the three couples (i.e., *rooster-voyage*, *cord-smile* and *brother-monk*) at time.

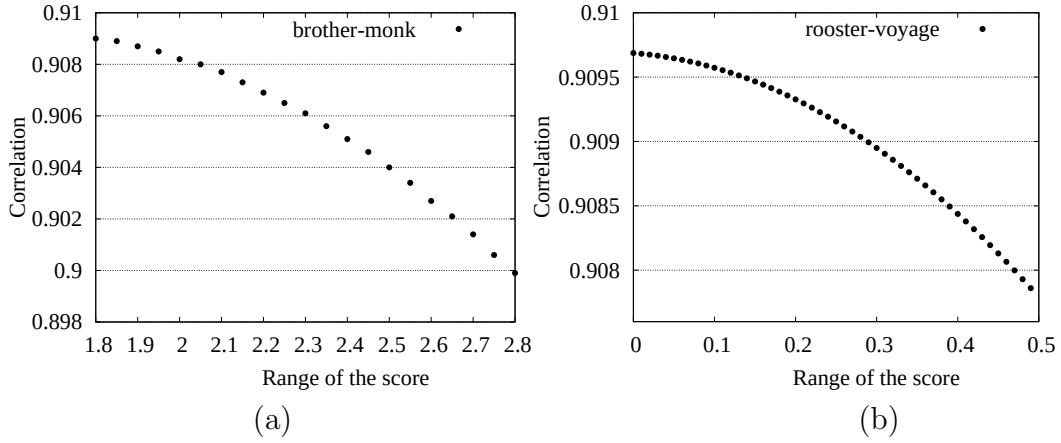


Fig. 6. Variation of the scores of *brother-monk* (a) and *rooster-voyage* (b).

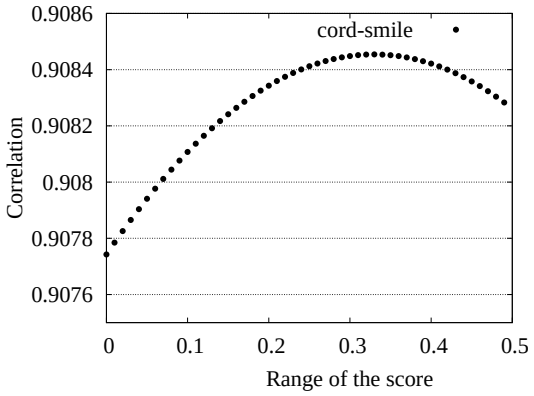
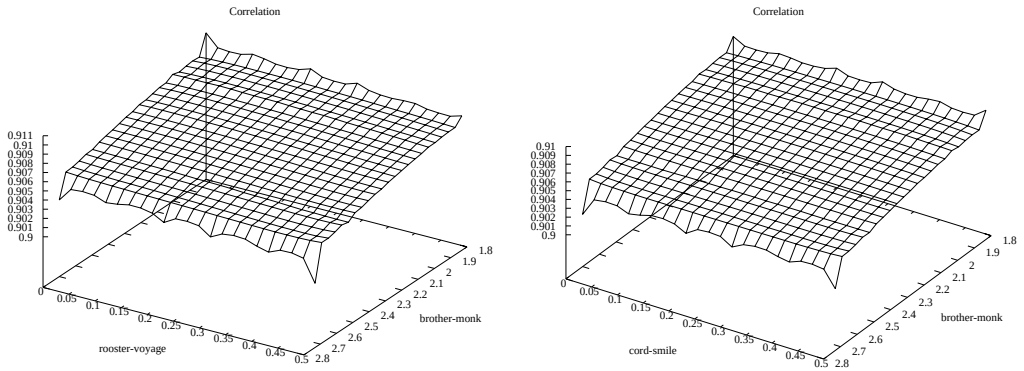


Fig. 7. Variation of the scores of *cord-smile*.

As can be observed, the correlation varies in the range 0.9-0.909 for the couple *brother-monk*, 0.907-0.910 for the couple *rooster-voyage* and 0.907-0.908 for the couple *cord-smile*. By recalling that the correlation of the P&S metric is 0.908 on the R&G dataset, we can observe that the lowest correlation (i.e., 0.9) is obtained when the score of *brother-monk* is lower than that given by the participants in the P&S experiment (i.e., 2.8 manual assigned vs. 1.997 obtained through the P&S experiment). However, this will bring a little worsening of the correlation.

Figures 8 and 9 report the values of correlation varying the scores of two pairs simultaneously. In these experiments we noted that the lowest value of correlations are 0.899 for the graphs reported in Fig. 8 (a) and Fig. 9 (b) whereas for the graph reported in Fig. 7 the lowest value of correlation is 0.907. Even in this case the value of correlation of the P&S metric is barely affected.



(a) (b)

Fig. 8. Variation of the scores of *brother-monk* for the couples *rooster-voyage* (a) and *cord-smile* and (b).

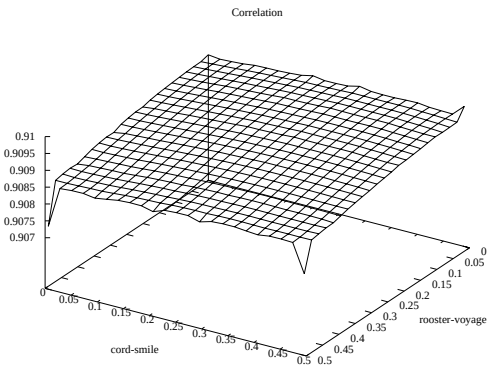


Fig. 9. Variation of the scores for the couples *rooster-voyage* and *cord-smile*.

Finally, we noted that varying the scores of the three couples simultaneously in the ranges shown in Table 10 the maximum value of correlation is 0.916 obtained with the scores of 0.3, 1.8 and 0 for *rooster-voyage*, *cord-smile* and *brother-monk* respectively. On the other hand, the lowest value of correlation is 0.898 obtained when the scores for the three pairs are 0, 2.8 and 0.5 respectively.

Overall, by analyzing the scores of the couples with the highest standard deviation we can conclude that the performance of the P&s metric would have been barely affected if these couples had received different (into the ranges reported in Table 10) scores.

5.5 Impact of Intrinsic Information Content

In Section 5.3 has been pointed out how in case the words to be compared are leaves, the Lin and J&C (transformed) metrics turn into the Resnik metric by adopting the intrinsic Ic formulation. Therefore, it is interesting to evaluate the impact of the intrinsic IC formulation on IC metrics. Fig. 10 shows the results of this evaluation. For sake of space we do not report the scores obtained by considering the two IC formulations. As can be noted, the correlation is improved for each metric. In particular, a notable improvement is reached by the J&C (about 40%) and P&S metrics (about 15%). That underlines how the intrinsic IC formulation is an effective and convenient way to compute IC values.

5.6 New Challenges for Researchers

The results obtained by some metrics in our experiments are very close to human judgments.

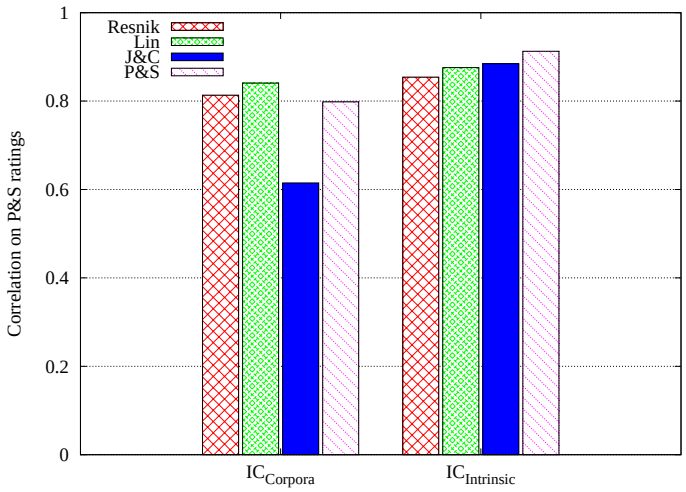


Fig. 10. Impact of the Intrinsic IC formulation

At this point a question arises: how much we can expect from a computational method for assessing semantic similarity ? Resnik in [24] took into account the correlation between experiments in order to obtain a possible upper bound. Resnik obtained a value of correlation w.r.t $M\&C$ experiment of 0.9583 while the inter-annotator agreement obtained was 0.9015. This latter result has been considered for many years as the theoretical upper bound. However, we agree with what was observed in [13] and propose to consider as upper bound not the inter-annotator agreement but the correlation between the rating of the different experiments. This is because semantic similarity should be considered as a collective property of groups of peoples (i.e., all the participant to the experiment) rather than considering them individually as done by Resnik with the inter-subject agreement. Moreover, since we replicated the $R\&G$ experiment on all the 65 word pairs dataset we can correlate our results with those obtained by $R\&G$. Hence, we propose to set as new hypothetical upper bound the value of correlation between the $R\&$G and $P\&S$ ratings, that is, 0.972. This latter consideration provides new challenges for researches. In fact, even if the metric we presented obtains a correlation value of 0.908 using this dataset, this value is far from the new hypothetical upper bound.

5.7 The Java WordNet Similarity Library (JWSL)

The P&S metric has been included in the Java WordNet Similarity Library (JWSL). JSWL[7], aims at providing developers with a library for accessing the WordNet lexical database. The main features of JWSL can be summarized

[7] The JWSL is available upon request. Refer to `http://grid.deis.unical.it/similarity` for further details

as follows: (i) it exploits a Lucene [8] index including the whole WordNet structure. This way, the computation of similarity between words can be speed up and does not require to install the WordNet software; (ii) it is written in Java. To the best of our knowledge, the most similar tool to JWSL is the WordNet::Similarity [9]. However, this valuable tool, is a web-based application (written in Perl). Another Java library, the JWNL [10] only provides access to WordNet and does not implement similarity metrics; (iii) it implements several similarity metrics also allowing new ones to be added.

6 Investigating the generality of the P&S metric

Most of current similarity metrics have been extensively evaluated on WordNet, which is a valuable source of general knowledge about the world. At this point it is valuable to investigate the generality of the approach we defined in terms of both the intrinsic formulation of IC and $P\&S$ similarity metric. To this aim, we performed a further evaluation by considering the MeSH ontology that, differently from WordNet, contains knowledge specific to a particular domain.

6.1 The MeSH Ontology

The MeSH Medical Subject Headings (MeSH) ontology is mainly a hierarchy of medical and biological terms defined by the U.S National Library of Medicine (NLM). It consists of a controlled vocabulary and a *Tree*. The controlled vocabulary contains several different types of terms such as *Descriptors*, *Qualifiers*, *Publication Types*, *Geographics* and *Entry* terms. Entry terms are the synonyms or the related terms to descriptors. For example, the descriptor *"Neoplasms"* has the following entry terms *"Cancer"*, *"Cancers"*, *"Neoplasm"*, *"Tumors"*, *"Tumor"*, *"Benign Neoplasm"*, *"Neoplasm"*, *"Benign"*. This vocabulary is used by the NLM to catalogue books, other library materials, and to index articles for inclusion in health related databases including MEDLINE. MeSH descriptors are organized in a tree which defines the MeSH Concept Hierarchy. In the MeSH tree there are 15 categories each of which is further divided into subcategories. For each subcategory, its descriptors are arranged in a hierarchy from most general to most specific [11].

[8] `http://lucene.apache.org`
[9] `http://wn-similarity.sourceforge.net`
[10] `http://wordnet.sourceforge.net`
[11] For an extensive discussion about the MesH structure and organization refer to `http://www.nlm.nih.gov/mesh`

6.2 Evaluation Methodology

As in the case of WordNet, in order to evaluate our metric it is necessary to have a set of human ratings. In the biomedical domain, there are no de facto standard human rating sets as the $P\&S$, $R\&G$ or $M\&C$ for WordNet. Thus, to evaluate our metric and compare it with state of the art, we adopt the set of 35 word pairs carefully elaborated and discussed in [7]. In this experiment, similarity was evaluated by doctors that gave a score to each pair between 0 (not similar) and 4 (perfect similarity). The average rating (by all doctors) of each pair represents an estimate of how similar each pair is according to human judgment.

6.3 Experimental Evaluation

Table 11 reports the similarity values provided by both humans and computational methods. Note that values obtained by the $P\&S$ metric have been normalized in the interval [0,1]. Moreover, for the Li metric we assigned the parameters α and β values of 0.2 and 0.6 respectively. Authors found that these parameter values bring the highest correlation w.r.t human judgment as discussed in [13]. As for the evaluation on WordNet, for IC based metrics the values of IC are obtained by the intrinsic formulation of IC. This approach will be useful to evaluate the domain independence of our formulation of IC. The correlation between computational methods and human judgments is reported in Table 12. The *p-value*, obtained as described in Section 5.1 even in this case is $p - value < 0.001$ which indicates that our results are highly significant.

6.3.1 Discussion of the Results

Interesting considerations can be done for the results reported in Table 12. The Li metric, which on WordNet was the the most close to the P&S metric, obtained the lowest correlation on MeSH. We hypothesize that this is mainly due to two reasons. The first concerns parameter values. The Li metric heavily depends on the parameters α and β to correctly balance the contribution of the path between the two concepts to be compared and the depth of their *msca*. Therefore, it is possible that parameter values that achieve a good correlation in a context cannot obtain the same (comparable) performance in another. The second reason is related to the structure of the considered ontology. MeSH is a more domain-specific ontology than WordNet and therefore, in MeSH the combination of path and depth in a non linear function as suggested by the Li metric could not have the same consistent interpretation as in WordNet. The three information content measures obtained better correlation.

Table 11
Similarity values obtained by the different metrics on Mesh

Word Pair			**IC-based**			**Hybrid**	**Features**
*Word*1	*Word*2	*Human*	*Resnik*	*Lin*	*J&C*	*Li*	*P&S*
Anemia	Appendicitis	0.031	0.000	0.000	0.190	0.130	0.133
Otitis Media	Infantile Colic	0.156	0.000	0.000	0.160	0.100	0.000
Dementia	Atopic Dermatitis	0.060	0.000	0.000	0.290	0.130	0.202
Bacterial Pneumonia	Malaria	0.156	0.000	0.000	0.030	0.100	0.000
Osteoporosis	Patent Ductus Arteriosus	0.156	0.000	0.000	0.150	0.000	0.000
Sequence	AntiBacterial Agents	0.155	0.000	0.000	0.270	0.160	0.184
Acq.Immunno. Syndrome	Congenital Heart Defects	0.060	0.000	0.000	0.070	0.080	0.000
Meningitis	Tricuspid Atresia	0.031	0.000	0.000	0.190	0.130	0.131
Sinusitis	Mental Retardation	0.031	0.000	0.000	0.360	0.130	0.117
Hypertension	Failure	0.500	0.000	0.000	0.210	0.130	0.109
Hyperlipidemia	Hyperkalemia	0.156	0.331	0.483	0.470	0.510	0.561
Hypothyroidism	Hyperthyroidism	0.406	0.619	0.726	0.750	0.630	0.718
Sarcoidosis	Tuberculosis	0.406	0.000	0.000	0.250	0.070	0.169
Vaccines	Immunity	0.593	0.000	0.000	0.520	0.000	0.344
Asthma	Pneumonia	0.375	0.517	0.790	0.870	0.520	0.749
Diabetic Nephropathy	Diabetes Mellitus	0.500	0.612	0.759	0.790	0.770	0.741
Lactose Intolerance	Irritable Bowel Syndrome	0.468	0.468	0.468	0.470	0.360	0.468
Urinary Tract Infection	Pyelonephritis	0.656	0.470	0.588	0.670	0.420	0.604
Neonatal Jaundice	Sepsis	0.187	0.000	0.000	0.190	0.160	0.000
Anemia	Deficiency Anemia	0.437	0.601	0.720	0.790	0.360	0.712
Psychology	Cognitive Science	0.593	0.627	0.770	0.810	0.800	0.751
Adenovirus	Rotavirus	0.437	0.267	0.332	0.450	0.350	0.398
Migraine	Headache	0.718	0.229	0.243	0.370	0.170	0.269
Myocardial Ischemia	Myocardial Infarction	0.750	0.595	0.918	0.890	0.800	0.830
Hepatitis B	Hepatitis C	0.562	0.649	0.823	0.860	0.660	0.790
Carcinoma	Neoplasm	0.750	0.246	0.626	0.850	0.450	0.651
Pulmonary Stenosis	Aortic Stenosis	0.531	0.658	0.781	0.810	0.660	0.763
Failure to Thrive	Malnutrition	0.625	0.000	0.000	0.180	0.130	0.126
Breast Feeding	Lactation	0.843	0.000	0.000	0.040	0.080	0.029
Antibiotics	Antibacterial Agents	0.937	1.000	1.000	1.000	0.990	1.000
Seizures	Convulsions	0.843	0.880	1.000	0.900	0.810	0.990
Pain	Ache	0.875	0.861	1.000	1.000	0.990	0.954
Malnutrition	Nutritional Deficiency	0.875	0.622	1.000	1.000	0.980	0.874
Measles	Rubeola	0.906	0.924	1.000	1.000	0.990	1.000
Chicken Pox	Varicella	0.968	1.000	1.000	1.000	0.990	1.000
Down Syndrome	Trisomy 21	0.875	1.000	1.000	1.000	0.990	1.000

Table 12
Correlation between computational methods and human judgments on MeSH

Metric	**Correlation**
Resnik	0.721
Lin	0.718
J&C	0.710
Li	0.707
P&S	0.725

Resnik's metric obtained a slightly higher level of correlation as compared to other IC-based metrics. This trend is in contrast with the results obtained by

the same metric on WordNet where it obtained the lower correlation both on the *M&C* and *R&G* datasets (see tables 8 and 9). This fact can be justified assuming the in MeSH the *msca* better expresses the amount of information shared by two terms. In this respect, it is important pointing out that for IC metrics the values of IC have been obtained by the intrinsic formulation discussed in Section 3.1. Finally, even on this dataset the *P&S* metric that exploits the feature-based theory of similarity and the intrinsic IC formulation obtained the highest correlation value. In this case the value of correlation is lower than that obtained on WordNet (i.e., 0.725 vs. 0.912). However, note that results have been proven to be significant due to the very low value of *p-value* (i.e., $p - value < 0.001$).

7 On extending the P&S metric

The P&S metric in its current implementation only considers the relations of *hypernymy/hyponomy* among concepts contained in a single ontology. However, it would be worth investigating how this metric can be extended in two main directions. The first one concerns cross-ontology similarity, that is, the problem of determining similarity between concepts belonging to two different ontologies. The second one is related to the other kinds of relations beyond *hypernymy/hyponomy* (e.g., *holonomy/meronymy*). These relations, in fact, can help refining the similarity (but in general the relatedness) between concepts.

An interesting definition of cross-ontology similarity is given in [26]. From this paper emerges that cross ontology similarity in an extension of single ontology similarity. Authors generalize the Tversky's model of similarity by including different components in the process of similarity computation. The proposed model of similarity takes as input two existing ontologies and connects them by adding a new virtual node (i.e., *anything*). This new node allows to compute an α coefficient, which expresses the "relative importance of the common and non-common characteristics". This coefficient is at the basis of the similarity measure, as it will be used by the sub-measures, and is computed by considering the depth of the two concepts to be compared. The different sub-measures to establish cross-ontology similarity consider synonym set, features and semantic-neighbors. Each of these sub-measures is based on the computation of the degree of intersection between the sets of synonym, features and semantic-neighbors and exploits the coefficient α. To compute the intersection, a string matching approach is exploited. The sub-measures are then combined using a weighting schema.

From this analysis emerged that in order to extend the P&S metric to compute cross-ontology similarity two main problems have to be addressed. The first one concerns the fact that it is not possible to find a *most specific common*

abstraction (msca) between the concepts to be compared. The msca is the common ancestor which, according to our formulation of similarity, expresses the degree of shared features. The second one is related to the values of intrinsic IC, which now should be computed considering concepts belonging to both ontologies. As for the first problem, we can adopt an approach similar to that described in [26] and consider the ontologies as connected by a new virtual concept. Hence, we can exploit informativeness of concepts (expressed by the intrinsic IC) as a factor for computing the amount of specific and shared featured between concepts.

Concerning the usage of a wider range of semantic relations, we can extend the notion of intrinsic information concepts by including other kind of relations in its computation. Therefore, the informativeness of a concept will take into account not only the number of its hyponyms but also the number of other relations such as part of relations.

8 Concluding Remarks

This paper presented a new similarity metric combining the feature based and information theoretic theories of similarity. We obtained the P&S metric by translating the Tversky formulation of similarity into the information theoretical domain. This metric, as shown by experimental evaluation, outperforms the state of the art. Moreover, the intrinsic IC formulation adopted in our metric, improves the results of other IC based metrics. Another contribution has been the similarity experiment we performed in order to build a reference basis for comparing the different metrics. This experiment involved a great number of participants and non-native English speakers, which is a novelty w.r.t previous experiments. We also made an interesting consideration about the upper bound of a computational method for assessing similarity thus providing new challenges for researchers. We implemented our metric and several others in the JWSL. Finally, to have an insight of the generality of both the metric and intrinsic IC formulation we performed a further evaluation of the MeSH ontology. Even in this case our metric obtained the best results of correlation w.r.t human judgments.

Acknowledgement

I would like to thank Nuno Seco for sharing his valuable insights during the many discussions we had about this topic. A special thank goes to prof. Domenico Talia, my PhD advisor, for his support. Finally I wish to thank the anonymous reviewers for their interesting remarks and suggestions that allowed to improve significantly the quality of the paper.

References

[1] R. L. Cilibrasi and P. M. B. Vitanyi, The Google Similarity Distance, *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370-383, 2007.

[2] F. M. Couto, M. J. Silva, and P. M. Coutinho, Measuring Semantic Similarity between Gene Ontology Terms, *Data & Knowledge Engineering*, 61(1):137-152, 2007.

[3] B. Danushka, M. Yutaka, and I. Mitsuru, Measuring Semantic Similarity between Words Using Web Search Engines, in: *Proceedings of WWW*, pp. 757-766, 2007.

[4] J. Devore, *Probability and Statistics for Engineering and the Sciences*, International Thomson Publishing Company, 1999.

[5] C. Hai, J. Hanhua, SemreX: Efficient Search in Semantic Overlay for Literature Retrieval, *Future Generation Computer Systems*, 24(6):475-488, 2008.

[6] G. Hirst and D. St-Onge, in *WordNet: An Electronic Lexical Database*, chapter Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms, MIT Press, 1998.

[7] A. Hliaoutakis, Semantic Similarity Measures in the MESH Ontology and their Application to Information Retrieval on Medline, Technical report, Technical Univ. of Crete (TUC), Dept. of Electronic and Computer Engineering, 2005.

[8] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios, Information Retrieval by Semantic Similarity, *International Journal on Semantic Web and Information Systems*, 2(3):55–73, 2006.

[9] K. Janowicz, Semantic Similarity Blog, http://www.similarity-blog.de.

[10] J. Jiang and D. Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, In *Proceedings of ROCLING X*, 1997.

[11] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, McGraw-Hill/Irwin, 2004.

[12] J. Lee, M. Kim, and Y. Lee, Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, *Journal of Documentation*, 49:188-207, 1993.

[13] Y. Li, A. Bandar, and D. McLean, An approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871-882, 2003.

[14] Y. Li, D. McLean, Z. Bandar, J. O'Shea, and K. Crockett, Sentence Similarity Based on Semantic Nets and Corpus Statistics, *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138-1150, 2006.

[15] D. Lin, An Information-Theoretic Definition of Similarity, In *Proceedings of Conference on Machine Learning*, pp. 296-304, 1998.

[16] C. Meilicke, H. Stuckenschmidt, and A. Tamilin, Repairing ontology mappings, In *Proceedings of AAAI*, pp. 1408-1413, 2007.

[17] G. Miller, Wordnet an On-line Lexical Database, *International Journal of Lexicography*, 3(4):235-312, 1990.

[18] G. Miller and W. Charles, Contextual Correlates of Semantic Similarity, *Language and Cognitive Processes*, 6(1):1-28, 1991.

[19] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, Measures of Semantic Similarity and Relatedness in the Biomedical Domain, *Journal of Biomedical Informatics*, 40(3):288-299, 2007.

[20] G. Pirró, and N. Seco, Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content, In *ODBASE*, pp. 1271-1288, 2008.

[21] G. Pirró, M. Ruffolo, and D. Talia, SECCO: On Building Semantic Links in Peer to Peer Networks. *Journal on Data Semantics*, XII: 1-36, 2009.

[22] R. Rada, H. Mili, and M. Bicknell, E. andBlettner, Development and Application of a Metric on Semantic Nets, *IEEE Transactions on Systems, Man, and Cybernetics*, 19:17-30, 1989.

[23] S. Ravi and M. Rada, Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity, In *Proceedings of ICSC*, 2007.

[24] P. Resnik, Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of IJCAI*, pp. 448-453, 1995.

[25] E. L. Rissland, AI and Similarity, *IEEE Intelligent Systems*, 21:39-49, 2006.

[26] M. Rodriguez and M. Egenhofer, Determining Semantic Similarity among Entity Classes from Different Ontologies, *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442-456, 2003.

[27] H. Rubenstein and J. B. Goodenough, Contextual Correlates of Synonymy, *Communications of the ACM*, 8(10):627-633, 1965.

[28] B. Schaeffer and R. Wallace, Semantic Similarity and the Comparison of Word Meanings, *J. Experiential Psychology*, 82:343-346, 1969.

[29] N. Seco, Computational Models of Similarity in Lexical Ontologies, Master's thesis, University College Dublin, 2005.

[30] N. Seco, T. Veale, and J. Hayes, An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of ECAI*, pp. 1089-1090, 2004.

[31] C. Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, 27:379-423, 1948.

[32] A. Tversky, Features of Similarity, *Psychological Review*, 84(2):327-352, 1977.

[33] A. Zavaracky, Glossary-based Semantic Similarity in the WordNet Ontology, Master's thesis, University College Dublin, 2003.

[34] V. S. Zuber and B. Faltings, OSS: A Semantic Similarity Function based on Hierarchical Ontologies. In *Proceedings of IJCAI*, pp. 551-556, 2007.