PAPER

Toward Simulating the Human Way of Comparing Concepts*

Raúl Ernesto MENÉNDEZ-MORA[†,††a)], ***Nonmember and*** **Ryutaro ICHISE**[†b)], ***Member***

SUMMARY An ability to assess similarity lies close to the core of cognition. Its understanding support the comprehension of human success in tasks like problem solving, categorization, memory retrieval, inductive reasoning, etc, and this is the main reason that it is a common research topic. In this paper, we introduce the idea of semantic differences and commonalities between words to the similarity computation process. Five new semantic similarity metrics are obtained after applying this scheme to traditional WordNet-based measures. We also combine the node based similarity measures with a corpus-independent way of computing the information content. In an experimental evaluation of our approach on two standard word pairs datasets, four of the measures outperformed their classical version, while the other performed as well as their unmodified counterparts.

key words: *WordNet, semantic similarity measures, information content, knowledge*

1. Introduction

An ability to assess similarity lies close to the core of cognition. Its understanding support the comprehension of human success in tasks like problem solving, categorization, memory retrieval, inductive reasoning, etc. In many fields such as artificial intelligence, biomedicine, linguistics, cognitive science, and psychology the semantic similarity of words is a topic of research. The computation of semantic similarity is extensively used in a variety of applications, like words sense disambiguation [1], detection and correction of malapropisms [2], information retrieval [3]–[5], automatic hypertext linking [6] and natural language processing. Several applications to the field of artificial intelligence are discussed in [7]. However, despite numerous practical applications today, its theoretical foundations lie elsewhere, in cognitive science and psychology where it has been the subject of many investigations and theories (e.g., [8]–[12]).

Let take a current example of peer-to-peer networks [13] into which semantic similarity has found its way. Assuming a shared taxonomy among the peers to which they can annotate their content, similarities among peers can be inferred by computing similarities among their representative concepts in the shared taxonomy. In this way, the more two peers are similar, the more efficient it is to route messages toward them. Numerous similar applications are the reasons for the increasing interest in this subject, whose ultimate goal is to mimic human judgment regarding similarity of word pairs.

Semantic similarity of words is often represented by the similarity between the concepts associated with the words. Several methods have been developed to compute word similarity, some of them operating on the taxonomic dictionary WordNet [14] and exploiting its hierarchical structure. However the majority of them suffer from a serious limitation. They only focus on the semantic information shared by those concepts, i.e., on the common points in the concept definitions or in the semantic differences but they never combine both. The increasing need for better measures and the new study area of semantic differences between words has led us to this study in the hope of upgrading existing semantic similarities measures. In particular, we combined traditional WordNet-based semantic similarity measures with the idea of the "similarity between entities being related to their commonalities as well as to their differences", in order to improve the performance of WordNet-based similarity measures and to obtain better results for applications using semantic similarities.

The paper is structured as follows. The next section reviews some background knowledge and related work. Section 3 describes our model, as well as the modified measures and the corpus-independent metric. Section 4 discusses the results of the experiment, and Sect. 5 summarizes our work, draws some conclusions, and outlines future work.

2. WordNet Similarity Measures

2.1 Semantic Similarity

Close to the core of cognition, similarity plays an indispensable foundational role in cognitive theories where several studies have been done. Four major psychological models of similarity are: geometric [8], featural [9], [10], alignment-based [11] and transformational [12].

Geometric models have been among the most influential approaches to analyzing similarity. Geometric models standardly assume minimality [$D(A, B) \geq D(A, A) = 0$], symmetry [$D(A, B) = D(B, A)$], and the triangle inequality [$D(A, B) + D(B, C) \geq D(A, C)$]. Tversky [10] criticized geometric models on the grounds that violations of

Manuscript received July 5, 2010.
Manuscript revised January 12, 2011.

[†]The authors are with National Institute of Informatics, Tokyo, 101–8430 Japan.
[††]The author is with Facultad de Informática y Matemática, Universidad de Holguín, Ave. XX Aniversario, Holguín, 80100 Cuba.

*This is an extended paper based on a research presented at IEA-AIE 2010

a) E-mail: menendez@nii.ac.jp, raul@facinf.uho.edu.cu
b) E-mail: ichise@nii.ac.jp

DOI: 10.1587/transinf.E94.D.1419

all three assumptions are empirically observed. When comparing things that are richly structured rather than just being a collection of coordinates or features often it is most efficient to represent things hierarchically (parts containing parts) and/or propositionally (relational predicates taking arguments). In such cases, comparing things involves not simply matching features, but determining which elements correspond to or align with one another. In alignment-based models, matching features influence similarity more if they belong to parts that are placed in correspondence, and parts tend to be placed in correspondence if they have many features in common and if they are consistent with other emerging correspondences. A fourth approach to modeling similarity is based on transformational distance. The similarity of two entities is assumed to be inversely proportional to the number of operations required to transform one entity so as to be identical to the other.

But the key to calculating semantic similarity lies in resembling human thinking behavior. Semantic similarity of concepts is determined by processing first-hand information sources in the human brain. Some studies have tried to assess the semantic proximity of two given concepts in order to improve the semantic similarity computation. These studies focus on similarity and they use synonymy[†], hyponymy[††] [15], meronymy[†††] and other arbitrarily typed semantic relationships among concepts. These relationships can be used to connect concepts in graph structures. They are the key ideas behind measures developed to assess the semantic similarity of concepts, i.e., how much one concept has to do with a different one. However, the measures tend to focus on the common points in the concepts' definitions; they rarely consider semantic differences, and in the best case both approach are never combined. This leaves a big gap in the semantic similarity computation process.

2.2 WordNet

A number of semantic similarity computation methods operate on the taxonomic dictionary WordNet [14] and exploit its hierarchical structure. WordNet is a machine-readable lexical database that is organized by meanings, and it was developed at Princeton University. Synonymy, hyponymy, meronymy and many other relationships between concepts are represented in this lexical network of English words. WordNet, as an ontology, is intended to model the human lexicon, and psycholinguistic findings were taken into account during its design. It is classified as a light-weight ontology, because it is heavily grounded on its taxonomic structure employing the IS-A inheritance relation, and as a lexical ontology, because it contains both linguistic and ontological information [16]. Figure 1 taken from [17] shows a fragment of WordNet's structure.

Nouns, verbs, adjectives, and adverbs are each organized into networks of synonym sets (*synsets*) each representing one underlying lexical concept and are interlinked with a variety of relations. A polysemous[††††] word will appear in one synset for each of its senses. The backbone of the noun network is the subsumption hierarchy (*hyponymy/hypernymy*), which accounts for close to 80% of the relations in WordNet.

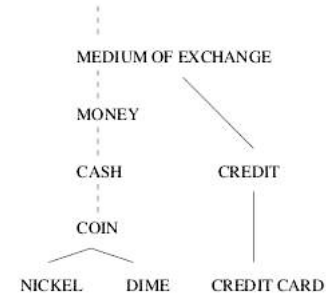

Fig. 1 Fragment of the WordNet taxonomy. Solid lines represent IS-A links; dashed lines indicate that some intervening nodes were omitted to save space.

2.3 Semantic Similarity Measures and WordNet

Based on WordNet and depending on the elements taken into consideration, semantic similarity measures can be classified into two different types: *edge-based similarity measures* and *node-based similarity measures*.

An intuitive way to quickly compute the semantic similarity between two nodes of a hierarchy is to count the number of edges in the shortest path between these two nodes. The idea behind this is that the semantic distance of two concepts is correlated with the length of the shortest path to join these concepts. This measure was first defined by Rada in [18]. However, it relies upon the assumption that each edge carries the same amount of information, which is not true in most ontologies [17]. Many other formulas have since extended Rada's measure by computing weights on edges by using additional information, such as the depth of each concept in the hierarchy and the *lowest common superset, or most specific subsumer (lcs)* [19]. For example, in Fig. 1, the *lcs* between the concepts *nickel* and *dime* is the concept *coin*.

The measures which focus on structural semantic information (i.e., the depth of the lowest common superset ($lcs(c_1, c_2)$), the depth of the concept's nodes, and the shortest path between them) are called *edge-based similarity measures*. In a paper on translating English verbs into Mandarin Chinese, Wu & Palmer [19] introduce a scaled metric for what they call *conceptual similarity* between a pair of concepts c_1 and c_2 in a hierarchy. Leacock & Chodorow [20] also rely on the length of the shortest path between two synsets for their measure of similarity. However they limit their attention to IS-A links and take into account the maxi-

[†]A semantic relation that holds between two words that can (in a given context) express the same meaning.

[††]The semantic relation of being subordinate or belonging to a lower rank or class.

[†††]The semantic relation that holds between a part and the whole.

[††††]Polysemy: The ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings.

Table 1 Compilation of the different similarity measures and their main features.

Type	Similarity	Description
Edge-based	Rada [18] Sim_{length}	Rely on the length of the shortest path joining two concepts.
	Wu & Palmer [19] Sim_{wup}	Rely on the depth of the lowest common superset between two concepts.
	Leacock & Chodorow [20] Sim_{lch}	Rely on the length of the shortest path between two synsets.
Node-based	Lin [23] Sim_{lin}	Defined by the ratio between the amount of information needed to state the commonality of the concepts and the information needed to fully describe what the concepts are.
	Resnik [17] Sim_{res}	Defined by the information content of the lowest common superset between two concepts.
	Pirró & Seco [24] $Sim_{P\&S}$	Based on Tversky's theory but from an information theoretic approach.
	Jiang & Conrath [22] $Sim_{j\&c}$	A combined approach where the edge counting scheme is enhanced by the information content approach.

mum depth of the taxonomy.

The Wu & Palmer [19] and Leacock & Chodorow [20] similarity measures are based in a linear model, whereas Li et al.'s approach [21] combines structural semantic information in a nonlinear model. Li et al.'s model empirically defines a similarity measure that uses the shortest path length, depth, and local density in a taxonomy. They include two parameters which represent the contribution of the shortest path length and the depth of the *lcs* in the similarity computation process.

Another way to compute the similarity between two nodes is by associating a weight with each node. Such similarity measures are called *node-based similarity measures*. The node-based similarity measures include the metrics of Resnik [17], Jiang & Conrath [22], Lin [23] and Pirró & Seco [24].

The first node-based similarity measure we will cover, was proposed by Resnik in [17]. It is defined by the information content (IC) of the *lowest common superset (lcs)* of concepts c_1 and c_2. Resnik's approach was the first one bringing together ontology and corpus. Many other propositions have been made after Resnik to combine the IC of the two target nodes and their *lcs* (e.g. [22], [23]). Both Lin's and Jiang & Conrath's formulation correct a problem with Resnik similarity metric; if one were to calculate $Sim_{res}(c_1, c_1)$[†]one would not obtain the maximal similarity. Jiang & Conrath [22] metric is a semantic distance measure which can be transformed to a similarity metric as shown in [7].

Lin [23] calculates semantic similarity using a formula derived from information theory. It uses the same elements as Jiang and Conrath, but in a different fashion. 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".

The Pirró & Seco [24] similarity metric is based on Tversky's theory [10] but from an information-theoretic perspective. This measure achieves very good results in the comparison to human judgments when it is combined with the notion of *intrinsic information content*. In the next section different approaches for computing the information content (IC) will be introduced. Table 1 shows a compilation of the different similarity measures and their main features.

$$Sim_{P\&S}(c_1, c_2) = \begin{cases} 3IC(lcs(c_1, c_2)) & \\ \quad -IC(c_1) - IC(c_2) & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases} \tag{1}$$

For a better understanding of the foundations of the model presented in this paper we also introduce Tversky's abstract featural model of similarity [10].

In 1977, Tversky presented a model named the *Contrast Model* which takes into account features that are common to two concepts and features specific to each. That is, the similarity of concept c_1 to 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. Admitting a function $\psi(c)$ that yields the set of features relevant to c, he proposed the following similarity function:

$$\begin{aligned} Sim_{tvr}(c_1, c_2) = {} & \alpha F(\psi(c_1) \cap \psi(c_2)) \\ & - \beta F(\psi(c_1)/\psi(c_2)) \\ & - \gamma F(\psi(c_2)/\psi(c_1)) \end{aligned} \tag{2}$$

where F is some function that reflects the salience of a set of features, and α, β and γ are parameters provided for differences in each component. According to Tversky, similarity is not symmetric, that is, $Sim_{tvr}(c_1, c_2) \neq Sim_{tvr}(c_2, c_1)$, because humans tend to focus more on one object than on the other depending on the way the relationship direction is taken into consideration during the comparison. For example, regarding the concept *dime* in Fig. 1, it is logical that one of it's most related concepts is *nickel*, but the same is not true in the opposite direction. The concept *nickel* is also like *gold, metal, etc.*

2.4 Information Content

Node-based similarity measures compute the similarity between two nodes by associating a weight with each node.

[†]From now on this notation $Sim_{abbr}(c_1, c_2)$ will be used for representing the similarity expression corresponding to the authors or model specified by *abbr* when comparing concepts c_1 and c_2.

From the perspective of information theory, this weight represents the *information content (IC)* of a concept. IC can be considered to be a measure that quantifies the amount of information a concept expresses. The more specialized a concept is, the heavier its weight will be.

The literature contains two main ways of computing information content. The most classical way is Resnik's approach with a corpus [17]:

$$IC(c) = -\log\ p(c) \tag{3}$$

where $p(c)$ is the probability of concept c in a corpus. Seco's approach [25] exploits the notion of *intrinsic information content* (IIC) which quantifies IC's values by scrutinizing how concepts are arranged in an ontological structure:

$$IIC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(\max_{wn})} \tag{4}$$

where *hypo* returns the total number of hyponyms of a given concept c and $\max_{wn}$ is a constant that indicates the total number of concepts in the corresponding WordNet taxonomy. This definition of IC enables obtaining IC values in a corpus-independent way. Despite all this previous work, WordNet-based semantic similarity measures still have problems, which we will discuss in the next section.

3. The Method

3.1 Menendez-Ichise Model

Most of the WordNet-based semantic similarity measures just take into consideration semantic commonalities among concepts for computing their values. The strength of semantic differences has been diminished or not fully exploited while their combination have been rarely considered from a broader perspective. Having all these elements in mind and considering the current structure of WordNet, we proposed the Menendez-Ichise model [26]. In this section, we introduce our model and its application to traditional WordNet based similarity metrics. The modifications to those metrics are founded on Tversky's Contrast Model theory of similarity [10] which is classified as a featural model of similarity.

Our model supports to be a specialization of Tversky's featured-based theory applied to traditional WordNet-based semantic similarity measures. Paraphrasing Tversky, we state that: "the similarity between two entities is related to their commonalities as well as to their differences", and our general model is described by the following expression:

$$Sim(c_1, c_2) = \alpha * Comm(c_1, c_2) - \beta * Diff(c_1, c_2) \tag{5}$$

where $Comm(c_1, c_2)$ stands for ***commonalities***, $Diff(c_1, c_2)$ for the ***differences***, and α and β are tuning factors ($0 \leq \alpha$) and ($0 \leq \beta$) that represent the importance of the commonalities and differences in the model. Because WordNet's structure is represented by an undirected graph we can't avoid assuming symmetry where there is none.

The use of semantic differences for computing semantic similarity and its combination with the semantic commonalities is a novel approach. In the next section, we explain how we applied our model to WordNet-based semantic similarity measures.

3.2 Semantic Commonalities and Differences in WordNet Based Metrics

The main features considered by WordNet-based similarity metrics are, the distance between nodes and the weight of the nodes. This in turn leads to two different approaches: *edge-based* and *node-based*, as mentioned above.

In our model, regardless of the approach used, we consider the information from the *root*[†] to the *lcs* as the *semantic commonalities* of the concepts c_1 and c_2; and the rest of the information from the *lcs* to each of the concepts c_1 and c_2 as the *semantic differences*. Hence, from the perspective of an edge-based approach, the differences are related to the shortest path between the two concepts while the commonalities are related to the depth of the *lcs*. In node-based approach, the differences are related to the information contained in the nodes representing the concepts but not contained in their *lcs*, because this last one its encapsulating the common information. For example, regarding the concepts *nickel* and *dime* in Fig. 1, the semantic commonalities are in their *lcs*, i.e, the taxonomy subgraph from the *root* to the $lcs(nickel, dime) = coin$. The semantic differences between both concepts is enclosed in the taxonomy subgraph from $lcs(nickel, dime)$ to both concepts but without considering any information from the *root* to the concept *coin*.

Equation (6) is a combination of the traditional length and depth of the *lcs* metrics, because each of them deal with the differences and the commonalities respectively. We consider the first term of Sim'_{length} [††] as the semantic commonalities between the concepts, which is twice the distance from the *root* to their *lcs*. The second term as the semantic differences in this case the distance between the two concepts.

$$Sim'_{length}(c_1, c_2) = \alpha * \left(1 - \frac{1}{2 * depth(lcs(c_1, c_2))}\right) - \beta * \left(1 - \frac{1}{length(c_1, c_2) + 1}\right) \tag{6}$$

$depth(c)$: depth of concept c in the taxonomy, where the depth of the most abstract node, the "root", is 1.

$length(c_1, c_2)$: number of edges from node c_1 to node c_2 in the taxonomy.

While Wu & Palmer measure relies on the depth of the lowest common superset between the concepts (*semantic commonalities*), the Leacock & Chodorow measure relies on the length of the shortest path between two synsets

[†] The most abstract node in the taxonomy.

[††] From now on this notation Sim'_{abbr} will be used for representing the modified similarity expression corresponding to the authors or model specified by *abbr*.

(*semantic differences*). Now for each of their modified expressions (Eq. (7) and Eq. (8)) we have considered both the *semantic differences* and the *semantic commonalities*; which were not taken into consideration in their original formulation. To follow the approach of their original expressions the commonalities and the differences have been normalized using a different approach for each case. While Wu & Palmer measure used a *normalization factor* (the addition of the concepts' depths in the taxonomy), Leacock & Chodorow metric, used the properties of the the logarithm function to soften its values after dividing by twice the taxonomy's depth.

$$Sim'_{wup}(c_1,c_2) = \alpha * \left(\frac{2 * depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)}\right) - \beta * \left(\frac{length(c_1,c_2)}{depth(c_1)+depth(c_2)}\right) \quad (7)$$

$$Sim'_{lch}(c_1,c_2) = \alpha * \left(-\log\left(\frac{depth(lcs(c_1,c_2))}{2*\lambda}\right)\right) - \beta * \left(-\log\left(\frac{length(c_1,c_2)}{2*\lambda}\right)\right)$$
$$\lambda : \text{maximum depth of the taxonomy.} \quad (8)$$

The modified Resnik's similarity measure $Sim'_{res}(c_1,c_2)$ considers the *semantic commonalities* to be the information content of the $lcs(c_1,c_2)$ and the *semantic differences* to be the information content encompassed by concepts, minus the one already considered in the $lcs(c_1,c_2)$.

$$Sim'_{res}(c_1,c_2) = \alpha * IC(lcs(c_1,c_2)) - \beta * (IC(c_1)+IC(c_2) - 2 * IC(lcs(c_1,c_2)) \quad (9)$$

After the application of our model, the modified Jiang & Conrath similarity expression $Sim'_{j\&c}(c_1,c_2)$ is identical to the one obtained for Resnik's measure, Eq. (9), and it is a generalization of the Pirró & Seco similarity measure, Eq. (1).

$$Sim_{P\&S} \subset Sim'_{Res}(c_1,c_2) = Sim'_{j\&c}(c_1,c_2) \quad (10)$$

According to Lin [23] "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". In Eq. (11), we add the *semantic differences* as the information content in each concept minus the one already considered in the $lcs(c_1,c_2)$ divided by the information needed to fully describe the concepts. For Lin's expression the information needed to fully describe the concepts becomes a *normalization factor* (see Table 2) whose effect we will discuss later.

$$Sim'_{lin}(c_1,c_2) = \alpha * \left(\frac{2 * IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_2)}\right) - \beta * \left(\frac{IC(c_1)+IC(c_2)-2*IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_2)}\right) \quad (11)$$

Table 2 Normalization factor used with different metric approaches.

Metric	Approach	Normalization Factor
Sim'_{wup}	edge-based	$depth(c_1)+depth(c_2)$
Sim'_{lin}	node-based	$IC(c_1)+IC(c_2)$

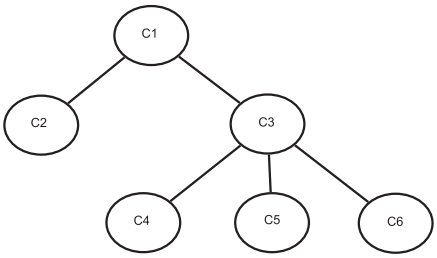

Fig. 2 Abstract taxonomy.

3.3 Extending the Intrinsic Information Content

Our model plan to take advantage of the benefits of our previous work [27], which extends Seco's intrinsic information content (IIC). From Seco's perspective, concepts that are leaf nodes are the most specific in the taxonomy so the information they express is maximal. This means it does not matter how deep is the leaf in the taxonomy. For example, in Fig. 2 concepts *C2* and *C4* should not have the same IC value, but since they both have the same number of hyponyms under Seco et al. approach they would have equal values of IC.

Although founded in the same ideas, Menendez and Ichise [27] extended Seco's model by considering the depth of the concept in the taxonomy ($depth(c)$) as an important factor. They stand for: *"The deeper a concept is found in a taxonomy means the amount of previous knowledge is larger and it should bear a higher value of information content"*.

We developed IC_{hd}, a corpus-independent information content metric [27], where the $_{hd}$ comes from the uses of the taxonomic properties number of hyponyms (h) and depth of the concept (d). However in Eq. (12) we introduced some slights modifications to our previous work. In this variation of IC_{hd} we added "+1" to the argument of the logarithm in the enumerator of the fraction to avoid undefined values of the *log* function when its argument is 0. To keep the uniformity we also increased in one unit the argument of the logarithm in the denominator of the fraction:

$$IC_{hd}(c) = 1 - \left(\frac{\log\left(hypo(c) * \left(max_{depth} - depth(c)\right)+1\right)}{\log\left(max_{wn} * max_{depth}+1\right)}\right) \quad (12)$$

where function $hypo(c)$ returns the number of hyponyms of a given concept, max_{wn} is a constant that is set to the maximum number of concepts in the taxonomy, function $depth(c)$ returns the depth of a given concept and max_{depth} represents the maximum depth of the corresponding taxonomy. When the number of hyponyms of a concept ($hypo(c)$)

Table 3 Values of the information content for various concepts using different IC approaches.

concept	pos	sense	hypo	depth	IC	IIC	IC_{hd}
entity	noun	1	74373	2	0.0	0.009	0.018
cock	noun	4	1	15	12.16	0.939	0.875
noon	noun	1	0	11	11.06	1.0	1.0

decrease, the fraction in the IC_{hd} expression tends to 0 and it moves IC_{hd} metric closer to its maximum value, 1. Similar behavior is observed when the concept is located deeper into the taxonomy. The difference between the maximum depth of the corresponding taxonomy and the concept's depth ($depth(c)$) moves closer to 0. When this difference is closer to 0 the fraction in the IC_{hd} expression tends to 0 and it moves IC_{hd} metric closer to its maximum value, 1.

Table 3 shows some examples of the information content value for various concepts using different IC computation approaches. The columns represent: the concept's string, the taxonomy where is located (noun, verb, adj, adv), the corresponding sense's number, the total number of hyponyms the concept have in the taxonomy, depth of the concept, the IC value (corpus-dependent), the IIC value (Seco's approach) and the IC_{hd} value (our approach).

4. Experiments and Results

4.1 Evaluation Procedure and Data

The purpose of the experiments is to prove the hypothesis that the use of semantics commonalities as well as semantics differences can improve the computation of similarity between concepts. We also want to test the effectiveness of the IC_{hd} corpora independent information content metric with node-based similarity measures. In the experiments we evaluate the new semantic similarity measures and establish a baseline for comparison of their results with those of their original versions.

Unfortunately, there is a distinct lack of standards for evaluating semantic similarities. Which means that the accuracy of a computational method for evaluating word similarity can only be established by comparing its results against human common sense. That is, a method that comes close to matching human judgments can be deemed accurate.

The Pearson correlation coefficient indicates the strength of a linear relationship between two variables. Although its value generally does not completely characterize their relationship, we will use it for comparing the results of our similarity measures and the human judgments. The Pearson correlation is +1 in the case of a perfect positive (increasing) linear relationship, −1 in the case of a perfect decreasing (negative) linear relationship [28], and some value between −1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

```
select dataset of word pairs;
for each similarity measure (Sim_i) {
  for each pair of concepts (c1_j,c2_j) in
      dataset {
    if (Sim_i is an edge-based similarity) {
      compute Sim_i(c1_j,c2_j);
    }
    else {
      // compute the similarity value using
      // different  information content
      // metrics (IC, IIC and IC_hd)

      for each information content metric
        compute Sim_i(c1_j,c2_j);
    }
  }

  //compute correlation factor
  Corr_Sim_i =
     Correlation_Factor(Sim_i,Human_Judgment);

  // compare Corr_Sim_i with the correlation
  // of the original similarit measure version
  compare(Corr_Sim_i, Corr_Orig_Version);
}
```

Fig. 3 Algorithm for evaluating the quality of the similarity measures.

The procedure for evaluating the quality of the developed similarity measures is described in Fig. 3. In general, after choosing the dataset of word pairs, we will compute the similarity between the words using different similarity measures. To deal with the polysemy property of words, the similarity for each possible combination of meaning for each word pair will be computed. And we will keep the pair of concepts whose similarity is maximal in the previous step. Each node-based similarity measures will be computed using three different information content metrics (IC, IIC, IC_{hd}). In all the measures the importance of the commonalities and the differences will be changed to assess the best ratio between them.

Some datasets of word pairs are commonly used for this evaluation. In particular, the Rubenstein and Goodenough dataset (R&G in the following) and the Miller and Charles dataset (M&C in the following) are standard datasets for evaluating semantic similarities.

In 1965, Rubenstein and Goodenough [29] obtained "synonymy judgments" of word pairs by hiring 51 subjects to evaluate 65 pairs of nouns. The subjects were asked to assign a similarity from 0 to 4, from "semantically unrelated" to "highly synonymous". Miller and Charles [30], 25 years later, extracted 30 pairs of nouns from the R&G dataset and repeated their experiment with 38 subjects. The M&C ex-

Table 4 Parameters used for the corpus-independent IC computation.

Parameter	Taxonomy	Value
$\max_{wn}$	Noun	82115
	Verb	25047
$\max_{depth}$	Noun	20
	Verb	14

periment achieved a correlation of 0.97 with the original experiment of R&G. Resnik [17], in 1995, replicated the M&C experiment with 10 computer science students, obtaining a correlation of 0.96. Pirró and Seco [24] in 2008 also recreated the R&G experiment this time with 101 subjects, and arrived at a correlation coefficient of 0.972 for the full dataset.

We used the human judgments of Pirró and Seco experiment [24] for the word pairs of both datasets, the Miller and Charles dataset (M&C) and the Rubenstein and Goodenough dataset (R&G). In M&C dataset we considered only 28 word pairs of the 30 used in the M&C experiment since a word missing in WordNet 3.0 made it impossible to compute ratings for the other two word pairs.

All the evaluations were performed using WordNet 3.0 [14] and the Brown Corpus[†] was used for the calculation of the corpus-dependent information content metric. For the computation we used Pedersen's WordNet::Similarity Perl module as the core. We also recreated the Pirró and Seco's experiment with the Java WordNet Similarity Library [24] (JWSL) using Pirró and Seco's intrinsic information content (IIC), but we did not obtain the same results they did. Probably due to the selection of the parameters. Table 4 shows the values we used for the parameters during the computation of the corpus independent information content metric (IIC and IC_{hd}) for the nouns and verbs WordNet's sub-taxonomies.

For the each metric we performed two experiments. The purpose of the first experiment is to check if the semantics differences have a positive effect in the performance of the measures when they are considered in the computation. In the second experiment we pursue to narrow the values for α and β which generate the higher performance of the semantic similarity measures. In both experiments, for the node-based measures we also check the effectiveness of the IC_{hd} information content.

In the first experiment we only variate the importance of the semantic differences' factor, β, and then we calculate the correlation of the new metrics' results with the human judgments values obtained in [24]. The importance of the semantic commonalities factor was kept constant ($\alpha = 1$) in this experiment, since we wanted to focus on the effect of semantic differences, but in the second experiment we do variate the importance of the semantics' commonalities as well. In both experiments we used an step of 0.10 for the variation of α and β. Each of the experiments was conducted with two different dataset of words' pairs, the M&C and the R&G datasets. Table 5 shows the general details for each experiment.

Table 5 Experiments description.

Experiment	α	β
Exp. 1	$\alpha = 1$	$\beta \geq 0$
Exp. 2	$\alpha > 0$	$\beta \geq 0$

4.2 Results and Discussion

Table 6 compiles the results of Exp. 1 for the edge-based similarity measures using several values for the *differences' factors* for the M&C dataset. Because of space limitation we did not include in the table all the values of β used to run the experiment, just the most representative. The second column, entitled "Original", represents the results of the original measure[††], i.e., the previous result. The correlation value for the unmodified functions and when $\beta = 0.0$ would be the same if the modified measure considers the commonalities as in the original metric. This is not the case for Sim'_{length} and Sim'_{lch} similarity measures and it is the reason for the difference in the correlation values between the original function and the modified version when $\beta = 0.0$. But going deeper in the details of the results we can say:

1. Sim'_{length} effectiveness improved when the semantic differences were considered (compared with path-length and depth of the *lcs* metrics). The ratio between the semantic differences and the semantic commonalities $\left(\frac{\beta}{\alpha} = 0.6\right)$ was 0.6 which suggests the importance of the commonalities is higher than the importance of the differences.

2. Sim'_{wup} effectiveness remains the same as its original version, showing no changes for any value of the semantic differences' importance, β. The normalization done for this measure is the reason of the unaltered results.

3. Sim'_{lch} slightly improved its correlation value when compared with its original expression since their absolute value is closer to 1. The correlation values of the modified version are negative confirming that our approach for considering the semantic commonalities and semantic differences is opposed to the approach used in the original measure. In Sim'_{lch} the semantic differences were considered a negative element in the expression while in the original formulation it was considered a positive element. Highly related or similar concepts obtained low values while unrelated concepts obtained high values. From this perspective the modified version is behaving like a distance rather than a similarity, therefore we could do the comparison using the absolute values of the correlation.

   For this similarity measure the ratio between the semantic differences and the semantic commonalities

[†]The Brown University Standard Corpus of Present-Day American English.

[††]The original metric have not been modified.

Table 6 Correlation coefficients obtained for edge-based measures in Exp. 1 using the 28 words' pairs of M&C dataset.

		β							
	Original	0.0	0.1	0.3	0.6	1.0	2.1	12	15
Sim'_{length}	0.8401	0.6673	0.7958	0.8498	**0.8571**	0.8549	0.8493	0.8421	0.8417
Sim'_{wup}	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**	**0.7726**
Sim'_{lch}	0.8293	−0.7126	−0.7446	−0.7804	−0.8039	−0.8165	−0.8261	**−0.8296**	−0.8295

Table 7 Correlation coefficients obtained for edge-based measures in Exp. 1 using the 65 words' pairs of R&G dataset.

		β							
	Original	0.0	0.1	0.3	0.6	1.0	2.1	12	15
Sim'_{length}	0.8373	0.4424	0.5974	0.7504	0.8200	0.8420	**0.8480**	0.8406	0.8400
Sim'_{wup}	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**	**0.7795**
Sim'_{lch}	0.8631	−0.6604	−0.7126	−0.7753	−0.8189	−0.8426	−0.8598	**−0.8644**	−0.8642

Table 8 Correlation coefficients obtained for node-based measures in Exp. 1 using the 28 words' pairs of M&C dataset and three different information content metrics.

			β						
		Original	0.0	0.3	0.6	1.0	2.1	2.8	4.5
Sim'_{lin}	IC	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**	**0.8587**
	IIC	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**	**0.8797**
	IC_{hd}	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**	**0.8821**
Sim'_{res}	IC	0.8308	0.8308	0.8555	0.8624	0.8655	0.8671	**0.8672**	0.8671
	IIC	0.8421	0.8421	0.8740	0.8816	0.8843	**0.8846**	0.8842	0.8833
	IC_{hd}	0.8361	0.8361	0.8743	0.8819	**0.8835**	0.8811	0.8796	0.8773
$Sim'_{j\&c}$	IC	−0.8660	0.8308	0.8555	0.8624	0.8655	0.8671	**0.8672**	0.8671
	IIC	−0.8805	0.8421	0.8740	0.8816	0.8843	**0.8846**	0.8842	0.8833
	IC_{hd}	−0.8712	0.8361	0.8743	0.8819	**0.8835**	0.8811	0.8796	0.8773

$\left(\frac{\beta}{\alpha} = 12\right)$ shows the semantic differences are more important than the commonalities for the similarity computation process.

Table 7 compiles the results of Exp. 1 for the edge-based similarity measures using a larger dataset of words' pairs, the R&G dataset. The general description for Table 6 are also valid in here but let go through the details:

1. Sim'_{length} effectiveness improved when the semantic differences were considered. The ratio between the semantic differences and the semantic commonalities $\left(\frac{\beta}{\alpha} = 2.1\right)$ was 2.1 which suggests the importance of the semantic differences is higher than the importance of the commonalities when we have a larger dataset like R&G.

2. Despite the different dataset, the modified expression of Wu & Palmer (Sim'_{wup}) remains the same as its original version showing no changes for any value of β.

3. Sim'_{lch} slightly improved its correlation value compared to the original version, since their absolute value is closer to 1, when the semantic differences were taken into consideration. Again the semantic differences seem to be more important than the commonalities.

Table 8 compiles the results of Exp. 1 for the node-based similarity measures using several *differences' factors* and three different information content metrics: IC, IIC and IC_{hd} for the M&C words' pairs dataset. Again the column entitled "Original" represents the results of the original measure. The correlation value for the unmodified expression and when $\beta = 0.0$ for $Sim'_{j\&c}$ measure are different because in $Sim'_{j\&c}$ the commonalities were not consider as in the original metric. After a deeper analysis of the results we can say:

1. Sim'_{lin} achieved its highest value when combined with the IC_{hd} metric. But a similar behavior to the one observed for Sim'_{wup} was showed by Sim'_{lin} which no matter the value of the importance factors for the semantics differences, it remains the same as its original version as result of the normalization.

2. Sim'_{res} measure obtained higher values of correlation than the original expression when the semantic differences were considered. The results of the measure when combined with IC_{hd} approach overcome the IC approach while remain as competitive as with the IIC approach. The ratio between differences and commonalities changed for IC, IIC and IC_{hd}. This similarity measure is an extension of $Sim_{P\&S}$. The correlation values for $Sim_{P\&S}$ using IC, IIC and IC_{hd} were (0.8655, 0.8843 and 0.8835) respectively. So, Sim'_{res} always behaved better than $Sim_{P\&S}$ with any information content criteria.

3. $Sim'_{j\&c}$ also obtained higher values than its original ex-

Table 9 Correlation coefficients obtained for node-based measures in Exp. 1 using the 65 words' pairs of R&G dataset and three different information content metrics.

			β						
		Original	0.0	0.3	0.6	1.0	2.1	2.8	4.5
Sim'_{lin}	IC	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**	**0.8812**
	IIC	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**	**0.8992**
	IC_{hd}	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**	**0.9007**
$Sim_{P\&S}$	IC	**0.8793**	**0.8793**	**0.8793**	**0.8793**	**0.8793**	**0.8793**	**0.8793**	**0.8793**
	IIC	**0.8944**	**0.8944**	**0.8944**	**0.8944**	**0.8944**	**0.8944**	**0.8944**	**0.8944**
	IC_{hd}	**0.8915**	**0.8915**	**0.8915**	**0.8915**	**0.8915**	**0.8915**	**0.8915**	**0.8915**
$Sim'_{j\&c}$	IC	−0.8689	0.8677	0.8792	**0.8802**	0.8792	0.8763	0.8750	0.8732
	IIC	−0.8848	0.8773	0.8928	**0.8949**	0.8944	0.8918	0.8907	0.8889
	IC_{hd}	−0.8747	0.8679	0.8903	**0.8928**	0.8915	0.8867	0.8847	0.8817

Table 10 Maximum values of correlation obtained for Sim'_{length} and Sim'_{lch} measures in Exp. 2 for M&C and R&G datasets.

	Pairs	Correlation		Parameters	
	Dataset	Original	Max.	α	β
Sim'_{length}	M&C	0.8401	**0.8571**	1.0	0.6
	R&G	0.8373	**0.8481**	0.5	0.9
Sim'_{lch}	M&C	0.8293	**−0.8296**	1.0	12
	R&G	0.8631	**−0.8647**	0.3	2.1

Table 11 Maximum values of correlation obtained for Sim'_{res} in Exp. 2 using different IC approaches for M&C and R&G datasets.

	Pairs	IC	Correlation		Parameters	
	Dataset	Metric	Original	Max	α	β
Sim'_{res}	M&C	IC	0.8308	**0.8672**	0.1	0.3
		IIC	0.8421	**0.8849**	0.2	0.3
		IC_{hd}	0.8361	**0.8835**	1.0	1.0
	R&G	IC	0.8677	**0.8803**	0.9	0.5
		IIC	0.8773	**0.8949**	0.9	0.6
		IC_{hd}	0.8679	**0.8928**	1.0	0.6

pression. The negative value of the original function is due to the reason that the original expression is a distance and not a similarity. Since the expression for $Sim'_{j\&c}$ is equal to Sim'_{res} the rest of the conclusions are the same as above.

Table 9 compiles the results of Exp. 1 for the node-based similarity measures using a larger dataset of words' pairs (R&G):

1. For this larger dataset Sim'_{lin} improved its effectiveness when combined with IC_{hd} approach but it was not affected by different importance of the semantic differences.

2. Sim'_{res} measure improved its effectiveness compared with its original expression when our model is applied. Again the combination with IC_{hd} approach overcome the IC approach while remain as competitive as with the IIC approach. For this larger dataset the ratio between the semantic differences and the semantic commonalities showed certain stability $\left(\frac{\beta}{\alpha} = 0.6\right)$, it was 0.6 for all the different information content approaches. Although this lead us to the idea the commonalities are more important than the differences. As we already say this similarity measure is an extension of $Sim_{P\&S}$. The correlation values for $Sim_{P\&S}$ using IC, IIC and IC_{hd} with R&G dataset were 0.8793, 0.8944 and 0.8915 respectively. So, again Sim'_{res} behaved better than $Sim_{P\&S}$ independently of the information content criteria.

3. $Sim'_{j\&c}$ achieved better results than its original expression.

The compilation of Exp. 2 for M&C and R&G datasets is shown in Table 10 and Table 11. We excluded Sim'_{wup} and Sim'_{lin} because α and β have not effect in their performance. As we showed in Eq. (10), Sim'_{res} is an extension of $Sim_{P\&S}$ and since its expression is equal to $Sim_{j\&c}$ their results will be the same, so we also omitted from the experiment. In the second experiment we also changed the values of α to obtain the highest correlation value of the measures.

Table 10 shows the results of Exp. 2 for Sim'_{length} and Sim'_{lch} measures from which we can arrive to the following conclusions:

1. For both datasets Sim'_{length} obtained higher values when our model is applied.

2. After the application of our model the Sim'_{lch} slightly improved its correlation values for both datasets. In both cases the ratio between the semantic differences and the semantic commonalities $\left(\frac{\beta}{\alpha}\right)$ showed higher importance for the semantic differences. This result its due to the original construction of the function where the differences had the leading vote.

Table 11 compiles the results of Exp. 2 for Sim'_{res} after the application of our model and when different information content approaches were applied. From this table we can arrive to the following conclusions:

1. Sim'_{res} always obtained better correlations values than its original formulation when our model is applied. The combination of the modified expression with the IC_{hd} approach improved the corpus-dependent metric while remain as competitive as with IIC metric. For the larger dataset (R&G) this measure showed some stability in the ration between the semantic differences and the semantic commonalities, close to the value $\frac{\beta}{\alpha} = 0.6$. This trend was not observed for the M&C dataset, probably due to the small number of pairs in the dataset.

Summarizing the results of Exp. 1 and Exp. 2 we can state the application of Menendez-Ichise model showed positive results for the Sim'_{length}, Sim'_{lch}, $Sim'_{j\&c}$ and Sim'_{res} measures which obtained higher values of correlation than their original expressions when the semantic differences between the concepts were taken into consideration. The use of the IC_{hd} approach for the node-based measures always showed better results than the corpus-dependent approach while remaining as competitive as the IIC metric, in the case of Sim'_{lin} it allowed to obtain the highest correlation value. Sim'_{wup} measure remain the same as its original version. All node-based similarity measures were superior to the edge-based ones. The experiments also suggested a larger dataset could be helpful for estimating a ratio of the importance between the semantic differences and the semantic commonalities.

5. Concluding Remarks and Future Work

In this paper we introduced new ideas in the computation of WordNet-based semantic similarity measures and we also extended a corpora independent approach for calculating the information content of a concept. The five new measures developed are modifications of traditional WordNet-based semantic similarity metrics. Supported by a featured-based theory, they incorporate the idea of semantic differences between concepts into the similarity computation process. The experimental results showed that, four of the measures outperformed their classical while the other measure performed the same as their classical versions. These results demonstrate the strengths and positive effects of including concepts semantic differences and the proposed information content metric during their semantic similarity computations. The extended corpora independent approach generated the highest value for one of the node-based measure, and in general it improved the results of the corpus-dependent model while remained as competitive as the intrinsic information content approach.

This research focus on WordNet-based semantic similarity measures. The studied similarity measures use the hyponymy relation, also known as the "is-a" relation, for computing the similarity between two concepts. Despite of the fact that about 80% of the relationships in the WordNet taxonomy are "is-a" relationships, it is a shortcoming of those measures not to consider other types of relations. The term "semantic relatedness" refers to several types of lexical relationships, including hyponymy/ hypernymy, synonymy, meronymy, antonymy, as well as any other unsystematic relationships, i.e. functional relationships. The application of our approach to semantic relatedness measures remains as an open area of research.

As future work, we would like to enlarge the words' pairs dataset. This could help us to estimate the ratio between semantic differences and semantic commonalities. We also would like to apply some machine learning methods to estimate the best ratio between differences and commonalities, and finally to apply our developed measures to a real problem.

References

[1] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," Artif. Intell. Res., vol.11, pp.95–130, 1999.
[2] A. Budanitsky and G. Hirst, "Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures," Proc. Workshop on WordNet and Other Lexical Resources, 2001.
[3] J.H. Lee, M.H. Kim, and Y.J. Le, "Information retrieval based on conceptual distance in is-a hierarchies," J. Doc., vol.49, no.2, pp.188–207, 1993.
[4] R.K. Srihari, Z. Zhang, A. Rao, H. Baird, and F. Chen, "Intelligent indexing and semantic retrieval of multimodal documents," Information Retrieval, vol.2, pp.245–275, 2000.
[5] A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis, and E.E. Milios, "Information retrieval by semantic similarity," Int. J. Semantic Web and Information Systems (IJSWIS), vol.2, no.3, pp.55–73, 2006.
[6] S.J. Green, "Building hypertext links by computing semantic similarity," IEEE Trans. Knowl. Data Eng. vol.11, no.5, pp.713–730, 1999.
[7] N. Seco, Computational models of similarity in lexical ontologies, Master's thesis, University College Dublin, 2005.
[8] W.S. Torgerson, "Multidimensional scaling of similarity," Psychometrika, vol.30, no.4, pp.379–393, 1965.
[9] L. Sjoberg, "A cognitive theory of similarity," Goteborg Psychological Reports, vol.2, no.10, 1972.
[10] A. Tversky, "Features of similarity," Psycological Review, vol.84, no.4, pp.327–352, 1977.
[11] D. Gentner, "Structure-mapping: A theoretical framework for analogy," Cognitive Science, vol.7, no.2, pp.155–170, 1983.
[12] S. Imai, "Pattern similarity and cognitive transformations," Acta Psychologica, vol.41, pp.433–447, 1977.
[13] J. Hai and C. Hanhua, "Semrex: Efficient search in a semantic overlay for literature retrieval," Future Gener. Comput. Syst., vol.11, no.6, pp.475–488, 2008.
[14] C. Fellbaum, ed., Wordnet: An Electronic Lexical Database, 1st ed., Bradford Books, 1998.
[15] C.N. Ziegler, K. Simon, and G. Lausen, "Automatic computation of semantic proximity using taxonomic knowledge," Proc. 15th ACM Int. Conf. Information and Knowledge Management, pp.465–474, 2006.
[16] L. Mazuel and N. Sabouret, "Semantic relatedness measure using object properties in an ontology," Proc. 7th Int. Semantic Web Conf., LNCS, vol.5318/2008, pp.681–694, 2008.
[17] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," Int. Joint Conf. Artificial Intelligence, vol.14, no.1, pp.448–453, 1995.
[18] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," IEEE Trans. Syst. Man Cybern., vol.19, no.1, pp.17–30, 1989.
[19] Z. Wu and M. Palmer, "Verb semantics and lexical selection," Proc. 32nd Annual Meeting of the Assoc. for Comp. Ling., pp.133–138, 1994.
[20] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," WordNet: A Lexical Reference System and its Application, pp.265–283, 1998.
[21] Y. Li, Z.A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," IEEE Trans. Knowl. Data Eng., vol.15, no.4, pp.871–882, 2003.
[22] J.J. Jiang and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," Proc. Int. Conf. Research in Comp. Ling., pp.19–33, 1997.
[23] D. Lin, "An information-theoretic definition of similarity," Proc.

15th Int. Conf. Machine Learning, pp.296–304, 1998.
[24] G. Pirró and N. Seco, "Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content," Proc. On the Move to Meaningful Internet Systems: OTM 2008, pp.1271–1288, 2008.
[25] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in wordnet," Proc. European Conf. Artificial Intelligence, pp.1089–1090, 2004.
[26] R.E. Menéndez-Mora and R. Ichise, "Effect of semantic differences in wordnet-based similarity measures," Proc. 23rd Int. Conf. Industrial, Engineering & Other Applications of Applied Intelligent Systems, pp.545–554, 2010.
[27] R.E. Menéndez-Mora and R. Ichise, "The role of taxonomy properties in information content metrics," Proc. Int. Symposium Matching and Meaning 2010, pp.22–26, 2010.
[28] S. Dowdy and S. Wearden, Statistics for Research, John Wiley & Sons, New York, 1983.
[29] H. Rubenstein and J.B. Goodenough, "Contextual correlates of synonymy," Commun. ACM, vol.8, no.10, pp.627–633, 1965.
[30] G. Miller and W. Charles, "Contextual correlates of semantic synonymy," Languages and Cognitive Processes, vol.6, no.1, pp.1–28, 1991.

Raúl Ernesto Menéndez-Mora received a M.Sc. degree in Applied Mathematics and Informatics for Management from University of Holguin, Cuba in 2006 and a second M.Sc. in Informatics from University of Granada, Spain in 2009. He is currently a Ph.D. candidate at National Institute of Informatics in Japan. His research interests include semantic web, machine learning and data mining.

Ryutaro Ichise received the Ph.D. degree in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 2000. From 2001 to 2002, he was a visiting scholar at Stanford University. He is currently an associate professor in principles informatics research division at National Institute of Informatics in Japan. His research interests include machine learning, semantic web and data mining.