ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-20

A generic framework for comparing semantic similarities on a subsumption hierarchy

Emmanuel Blanchard[1] and **Mounira Harzallah**[1] and **Pascale Kuntz**[1]

Abstract. Defining a suitable semantic similarity between concept pairs of a subsumption hierarchy is becoming a generic problem for many applications in knowledge engineering exploiting ontologies. In this paper, we define a generic framework which can guide the proposition of new measures by making explicit the information on the ontology which has not been integrated into existing definitions yet. Moreover, this framework allows us to rewrite numerous measures, originally proposed in various contexts, which are in fact closely related to each other. From this observation, we show some metrical and ordinal properties. Experimental comparisons on WordNet and on collections of human judgments complete the theoretical results and confirm the relevance of our propositions.

1 Introduction

Semantic similarity is a generic issue in a variety of applications in the areas of computational linguistics, artificial intelligence and biology, both in the academic community and the industry. Examples include word sense disambiguation [20], detection and correction of word spelling errors (malaproprisms) [4], image retrieval [23], information retrieval [13] and biological issues [25].

Similarities have been widely studied for set representations. The similarity $\sigma(A, B)$ between two subsets of elements A and B is often defined as a function of the elements common to A and B and as a function of the distinct ones. The Jaccard's coefficient [12] and the Dice's coefficient [7], which have originally been defined for ecological studies, are probably the most commonly used similarities among a large family of coefficients [11][24]. Their theoretical properties have been carefully studied [10][6].

Another important issue is the evaluation of semantic similarity in a network structure. With a long history in psychology [27][21], the problem of evaluating semantic similarity in a network structure has known a noticeable renewed interest linked to the development of the semantic web. In the 1970's many studies on categorization were influenced by a theory which stated that, from an external point of view, the categories in a set of objects were organized in a taxonomy according to an abstraction process. It is a common principle of the current knowledge representation systems to describe proximity relationships between domain concepts by a hierarchy, or more generally by a graph, i.e. by the ontologies associated with the new languages of the semantic Web –in particular OWL [1].

The tree-based similarities defined on a subsumption hierarchy contain two categories of similarities: those which, like the Wu and Palmer's similarity [28], only depend on the hierarchical structure (e.g., path lengths between concept pairs), and those which, like the Lin's similarity [14], additionally incorporate statistics on a corpus (e.g., concept occurrence frequencies). Some recent work has tried to extend the tree-based definitions to graphs by simultaneously taking into account different semantic relationships [15]. But, despite its pertinence, this attempt is faced with many open problems, and in practice the set-based and the tree-based similarities still remain the most widely used.

Our main purpose here is to show that these measures, which have originally been proposed in various contexts, are closely related to each other. Most set-based similarities $\sigma(A, B)$ can be re-written as functions $f(|A|, |B|, |A \cap B|)$ of the cardinalities of sets A and B and of their intersection set $A \cap B$. In data analysis, a classification attempt, not widely used in knowledge engineering, has permitted to gather numerous similarity definitions into two parametrized functions that we denote by f_α and f_β [6]. In this paper, we extend the definitions of these functions to the tree-based similarities: we define two generic functions $\widetilde{f}_\alpha$ and $\widetilde{f}_\beta$ with the same schema as f_α and f_β. Each function depends on a real parameter α or β, and on the "information content" $\psi(c_i) = -\log P(c_i)$ initially introduced by Resnik [19], where $P(c_i)$ is the probability of encountering an instance of the concept c_i. The operational computation of the theoretical probability $\mathrm{P}(c_i)$ may vary according to the available information (e.g., a corpus). We show that numerous published tree-based similarities are associated with a α or β value and an approximation of P.

The interests of this work are threefold. First, some partial pairwise comparisons have already been presented in the literature, but our unified framework allows to precisely identify the theoretical differences and commonalities of a large set of measures. Second, an analysis of the combinatorics of the subsumption hierarchy has led us to define new approximations of the probability P which exploit information on the subsumption hierarchy which has not been integrated into existing measures yet. Third, we show that ordinal and metrical properties can be straightforwardly deduced from this unified framework.

We complete this theoretical study by numerical experiments on WordNet samples (version 2.0) and on benchmarks on which human judgments have been collected.

2 A typology of set-based similarities

In this section, we denote by S a finite set of elements and A, B, C some subsets of S. We briefly recall that a similarity σ on $\mathcal{P}(S)$ is a function $\sigma : \mathcal{P}(S) \times \mathcal{P}(S) \rightarrow \mathbb{R}_+$ which satisfies two properties: symmetry ($\sigma(A, B) = \sigma(B, A)$) and maximality ($\sigma(A, A) \geq \sigma(B, C)$). Most of the set-based similarities can be grouped into two parametrized families.

The first one σ_α has been proposed by Caillez and Kuntz [6]. It is defined by a ratio between the cardinality of the intersection $|A \cap B|$

[1] University of Nantes, France, email: name.surname@univ-nantes.fr

and the Cauchy's mean [5] of the cardinalities of the respective sets $|A|$ and $|B|$:

$$\sigma_\alpha(A,B) = f_\alpha(|A|,|B|,|A\cap B|) = \frac{|A\cap B|}{\mu_\alpha(|A|,|B|)} \quad (1)$$

where $\mu_\alpha(|A|,|B|) = \left(\frac{|A|^\alpha+|B|^\alpha}{2}\right)^{\frac{1}{\alpha}}$ for $\alpha \in \mathbb{R}$.

Note that the case $\alpha = 1$ concides with the classical arithmetic mean. The second family σ_β has been studied by Gower and Legendre [10]:

$$\sigma_\beta(c_i,c_j) = f_\beta(|A|,|B|,|A\cap B|) = \frac{\beta\cdot|A\cap B|}{|A|+|B|+(\beta-2)\cdot|A\cap B|} \quad (2)$$

where $\beta \in \mathbb{R}_+^*$.

Table 1 shows the correspondence for different values of α and β with well-known measures (see [24] for the original references of the definitions).

Table 1. Correspondence between different parameter values and well-known set-based similarities

α	Mean μ_α	Similarity σ_α
$-\infty$	minimum	Simpson
-1	harmonic	Kulczinsky
0	geometric	Ochiaï
1	arithmetic	Dice
$+\infty$	maximum	Braun&Blanquet

β	Similarity σ_β
1/2	Sokal&Sneath
1	Jaccard
2	Dice

It is easy to check that the values of the similarities σ_α and σ_β are in the interval $[0;1]$.

3 A new formulation of tree-based similarities

In the following, we denote by $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ a finite set of concepts. Formally, an ontology can be modeled by a directed graph where the nodes represent concepts and the arcs represent labeled relationships. Here, like often in the literature, we restrict ourselves to the subsumption relationship "is-a" on $\mathcal{C}\times\mathcal{C}$. This relationship is common to every ontology, and different papers have confirmed that it is the most structuring one (e.g., [18]). In this case, if we assume that each concept c_i has no more than one parent (direct subsumer), the ontology can be modeled by a rooted tree $T(\mathcal{C})$ where the root c_0 is either an informative concept or a "dummy" concept just added for the connectivity. We denote by c_{ij} the most specific common subsumer of the concepts c_i and c_j in $T(\mathcal{C})$.

In this section, we adapt the definitions 1 and 2 above to define new tree-based similarity families using the information content notion [19]. We also propose different ways to compute the information content of a concept which aims at better exploiting the hierarchy. Moreover, we show how our framework support the rediscovering of existing tree-based similarities. Our proposition allows to better understand both the relationships between the set-based similarities and the tree-based similarities and between the tree-based similarities themselves.

3.1 Two new generic functions

Like Lin in his seminal paper [14], let us suppose that a concept c_i references a subset $\mathcal{I}_i$ of an instance set $\mathcal{I}$. By analogy with the Shannon's information theory, the information content of the concept c_i is measured by $\psi(c_i) = -\log \mathrm{P}(c_i)$ where $\mathrm{P}(c_i) \in [0,1]$ is the probability for a generic instance of c_i to belong to $\mathcal{I}_i$. Similarly, the common information associated with a concept pair $\{c_i, c_j\}$ is the information content $\psi(c_{ij}) = -\log \mathrm{P}(c_{ij})$ of their most specific common subsumer c_{ij}.

Consequently, from the definitions 1 and 2, we deduce two new parametrized functions which define tree-based similarities:

$$\widetilde{\sigma}_\alpha(c_i,c_j) = \widetilde{f}_\alpha(\psi(c_i),\psi(c_j),\psi(c_{ij})) = \frac{\psi(c_{ij})}{\mu_\alpha\big(\psi(c_i),\psi(c_j)\big)} \quad (3)$$

where μ_α is the Cauchy's mean and $\alpha \in \mathbb{R}$, and

$$\begin{aligned}\widetilde{\sigma}_\beta(c_i,c_j) &= \widetilde{f}_\beta(\psi(c_i),\psi(c_j),\psi(c_{ij})) \\ &= \frac{\beta\cdot\psi(c_{ij})}{\psi(c_i)+\psi(c_j)+(\beta-2)\cdot\psi(c_{ij})}\end{aligned} \quad (4)$$

where $\beta \in \mathbb{R}_+^*$

Let us remark that $\widetilde{\sigma}_\alpha(c_i,c_j) = \widetilde{\sigma}_\beta(c_i,c_j)$ when $\alpha = 1$ and $\beta = 2$. The parameter α allows to choose different definitions of the mean (e.g., arithmetic, geometric). Formulation 4 explicitely shows that the parameter β allows to weight the importance of the common information associated with the most specific common subsumer. The logarithm base has no influence over this similarity measure due to the use of a ratio.

3.2 Information content computation

Let us remark that in practice the instance set $\mathcal{I}$ is never completely described in extension. Consequently, the operational computation of the probability $P(c_i)$ depends both on the information at our disposal and on the hypothesis carried through the construction of the ontology. We denote by $\widehat{P}(c_i)$ the approximation of $P(c_i)$ in practice.

The approximation $\widehat{\mathrm{P}}_r$ proposed by Resnik is computed by the formula: $\widehat{\mathrm{P}}_r(c_i) = \frac{n(c_i)}{n(c_0)}$ where $n(c_i)$ is the number of occurrences of c_i plus the number of occurrences of the concepts which are subsumed by c_i in $T(\mathcal{C})$. This approximation considers the root as virtual ($\widehat{\mathrm{P}}_r(c_0) = 1$).

The probability $\mathrm{P}(c_i)$ can be approximated without considering any additional information. We propose some approximations deduced from various hypothesis on the extension of the concepts. We distinguish three approaches associated with different hypothesis:

- descending approach
  - Hypothesis 1: exponential decreasing of the instance number with concept depth in $T(\mathcal{C})$ ($\widehat{\mathrm{P}}_d$)
  - Hypothesis 2: uniform distribution of the father's instances on its sons ($\widehat{\mathrm{P}}_s$)
- ascending approach
  - Hypothesis 3: exponential increasing of the instance number with concept height in $T(\mathcal{C})$ ($\widehat{\mathrm{P}}_h$)
  - Hypothesis 4: uniform distribution of the root's instances on leaves ($\widehat{\mathrm{P}}_g$)
- combined approach
  - $\widehat{\mathrm{P}}_{dh}$: aggregation of $\widehat{\mathrm{P}}_d$ and $\widehat{\mathrm{P}}_h$
  - $\widehat{\mathrm{P}}_{sg}$: aggregation of $\widehat{\mathrm{P}}_s$ and $\widehat{\mathrm{P}}_g$

3.2.1 *Approximation $\widehat{\mathrm{P}}_d$ (Hypothesis 1)*

The probability for an instance to be associated with a concept c_i decreases exponentially with the depth d_i of c_i in $T(\mathcal{C})$. Then,

$$\widehat{\mathrm{P}}_d(c_i) = \frac{\widehat{\mathrm{P}}_d(parent(c_i))}{k} = \frac{\widehat{\mathrm{P}}(c_0)}{k^{d_i}} \quad (5)$$

where k is a fixed integer and $parent\,(c_i)$ is the parent (direct subsumer) of c_i.

Let us remark that when the logarithm base is set to k, the information content of a concept c_i is equivalent to its depth plus the information content of the root:

$$\psi_d(c_i) = -\log_k \widehat{\mathrm{P}}_d(c_i) = d_i + \psi(c_0) \tag{6}$$

3.2.2 Approximation $\widehat{\mathrm{P}}_s$ (Hypothesis 2)

We consider a uniform distribution of the instances of a father concept on its son concepts :

$$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(parent(c_i))}{|Children(parent(c_i))|} \tag{7}$$

where $Children\,(c_i)$ corresponds to the set of sons of c_i.

The information content (ψ_s) deduced from this approximation corresponds to the specificity degree in comparison with the root ; the depth takes into account a part of the information exploited by this specificity degree. This approximation refines $\widehat{\mathrm{P}}_d$ by considering the number of sons of each subsumer.

3.2.3 Approximation $\widehat{\mathrm{P}}_h$ (Hypothesis 3)

Each leaf has the same instance number and the probability of an instance to be associated with a concept c_i increases exponentially with the height of c_i. A leaf concept has a minimal probability which depends on the height of the hierarchy and on the instance number of the root. We can approximate $\mathrm{P}(c_i)$ by:

$$\widehat{\mathrm{P}}_h(c_i) = \frac{\widehat{\mathrm{P}}(c_0)}{k^{h_0 - h_i}} \tag{8}$$

In the particular case of a logarithm base equal to k, the information content of a concept c_i is defined by:

$$\psi_h(c_i) = -\log_k \widehat{\mathrm{P}}_h(c_i) = h_0 - h_i + \psi(c_0) \tag{9}$$

3.2.4 Approximation $\widehat{\mathrm{P}}_g$ (Hypothesis 4)

We consider a uniform distribution of the instances of the root concept on the leaf concepts:

$$\widehat{\mathrm{P}}_g(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{|Leaves(c_i)|}{|Leaves(c_0)|} \tag{10}$$

where $Leaves\,(c_i)$ corresponds to the leaf set subsumed by c_i (when c_i is a leaf, $Leaves\,(c_i) = \{c_i\}$).

This case is dual to the previous $\widehat{\mathrm{P}}_s$ case. Here, the information content (ψ_g) deduced from this approximation corresponds to the generality degree in comparison with the leaves ; the height takes into account a part of the information exploited by this generality degree. This approximation refines $\widehat{\mathrm{P}}_h$ by considering the number of sons of the concept and its subsumed concepts.

3.2.5 Approximations $\widehat{\mathrm{P}}_{sg}$ and $\widehat{\mathrm{P}}_{dh}$

We consider an alternative which simultaneously take into account the specificity and the generality degrees:

$$\widehat{\mathrm{P}}_{sg}(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i) + \widehat{\mathrm{P}}_g(c_i)}{2} \tag{11}$$

The definition of $\widehat{\mathrm{P}}_{sg}$ is based on the arithmetic mean of $\widehat{\mathrm{P}}_s$ and $\widehat{\mathrm{P}}_g$. This choice is forced by the preservation of the recursivity: $\widehat{\mathrm{P}}_{sg}(c_i) = \sum_{c_x \in Children(c_i)} \widehat{\mathrm{P}}_{sg}(c_x)$.

A dual case is the aggregation of $\widehat{\mathrm{P}}_d$ and $\widehat{\mathrm{P}}_h$:

$$\widehat{\mathrm{P}}_{dh}(c_i) = \frac{\widehat{\mathrm{P}}_d(c_i) + \widehat{\mathrm{P}}_h(c_i)}{2} \tag{12}$$

3.3 Similarity definitions deduced from the approximations

In this subsection, we show that the generic functions $\widetilde{\sigma}_\alpha$ and $\widetilde{\sigma}_\beta$ describe a set of semantic similarities (e.g., Lin, Wu & Palmer). We show that, in some cases, the approximations of $P\,(c_i)$ coincide with known measures of the literature.

3.3.1 Lin's similarity

The Lin's similarity [14] is analogous to the Dice's coefficient with the Resnik's approximation:

$$lin(c_i, c_j) = \frac{2 \cdot \psi_r(c_{ij})}{\psi_r(c_i) + \psi_r(c_j)} \tag{13}$$

Due to the Resnik's approximation, the root concept is considered as virtual ($\widehat{\mathrm{P}}(c_0) = 1$).

3.3.2 Wu & Palmer's similarity

The Wu & Palmer's similarity [28] is analogous to the Dice's coefficient with the approximation $\widehat{\mathrm{P}}_d$:

$$wup(c_i, c_j) = \frac{2 \cdot \psi_d(c_{ij})}{\psi_d(c_i) + \psi_d(c_j)} \tag{14}$$

3.3.3 Stojanovic's similarity

The approximation $\widehat{\mathrm{P}}_d$ allows to rewrite the Stojanovic's similarity [26] which is analogous to the Jaccard's coefficient:

$$sto(c_i, c_j) = \frac{\psi_d(c_{ij})}{\psi_d(c_i) + \psi_d(c_j) - \psi_d(c_{ij})} \tag{15}$$

3.3.4 Proportion of Shared Specificity

The Proportion of Shared Specificity (pss) proposed by Blanchard et al. [2] coincides with the Dice's coefficient with the $\widehat{\mathrm{P}}_s$ approximation:

$$pss(c_i, c_j) = \frac{2 \cdot \psi_s(c_{ij})}{\psi_s(c_i) + \psi_s(c_j)} \tag{16}$$

4 Metrical and ordinal properties

Most of the work on the mathematical properties of the similarities are focused on their metrical aspect [18]. They usually resort to preliminary transformations of the similarity into a dissimilarity of the form $\delta = Max_\sigma - \sigma$, where Max_σ is the maximal value reached by σ, or $\delta = \frac{1}{\sigma}$ when Max_σ is not finite, in order to check the triangular inequality $\delta\,(c_i, c_j) \leq \delta\,(c_i, c_k) + \delta\,(c_k, c_j)$.

Here, $Max_{\sigma_\alpha} = Max_{\sigma_\beta} = 1$ and we can consider the transformations $\delta_\alpha = 1 - \sigma_\alpha$ and $\delta_\beta = 1 - \sigma_\beta$. By studying the set-based similarities, Caillez et al. [6] and Gower et al. [10] have proved that the triangular inequality holds for $\alpha \to +\infty$ and $\beta \in [0, 1]$.

From a formal point of view, these questions are interesting; however, for practical applications in knowledge engineering, the developed approaches do not generally require this constraining property. When comparing results with different similarities, we can remark that specialists are more often concerned with the ordering associated with the obtained values than with the intrinsic values. Indeed, they order the concept pairs according to the proximities quantified by these measures.

Proposition 1. The similarities of the family $\{\widetilde{\sigma}_\beta\}_{\beta\in\mathbb{R}_+^*}$ follow the same ordering: for any c_i, c_j, c_k, c_l in $\mathcal{C}$, $\widetilde{\sigma}_\beta(c_i,c_j) \leq \widetilde{\sigma}_\beta(c_k,c_l) \Leftrightarrow \widetilde{\sigma}_{\beta'}(c_i,c_j) \leq \widetilde{\sigma}_{\beta'}(c_k,c_l)$ for any β and $\beta' \in \mathbb{R}_+^*$.

We show that $\widetilde{\sigma}_\beta(c_i,c_j) \leq \widetilde{\sigma}_\beta(c_k,c_l) \iff \widetilde{\sigma}_1(c_i,c_j) \leq \widetilde{\sigma}_1(c_k,c_l)$ for any $\beta \in \mathbb{R}_+^*$. When $\psi(c_i)+\psi(c_j)-2\cdot\psi(c_{ij})=0$ then, $\widetilde{\sigma}_1(c_i,c_j)=\widetilde{\sigma}_\beta(c_i,c_j)$ for any $\beta>0$. Otherwise, it is easy to check that, for $\psi(c_i)+\psi(c_j)-2\cdot\psi(c_{ij})\neq 0$,

$$\widetilde{\sigma}_\beta(c_i,c_j) = \frac{\beta\cdot\widetilde{\sigma}_1(c_i,c_j)}{1+(\beta-1)\cdot\widetilde{\sigma}_1(c_i,c_j)}$$

Consequently, $\widetilde{\sigma}_1(c_i,c_j) \geq \widetilde{\sigma}_1(c_k,c_l) \iff \widetilde{\sigma}_\beta(c_i,c_j) \geq \widetilde{\sigma}_\beta(c_k,c_l)$.

Proposition 2. The similarities of the family $\{\widetilde{\sigma}_\alpha\}_{\alpha\in\mathbb{R}}$ do not follow the same ordering.

Let us consider the following counter-example on a set $\mathcal{C} = \{c_1,c_2,c_3,c_4\}$. We suppose that c_1 is a subsumer of c_2, and that $\psi(c_1)=1$, $\psi(c_2)=3$, $\psi(c_3)=\psi(c_4)=2$ and $\psi(c_{34})=2$. In this case, the Cauchy's means are $\mu_\alpha(\psi(c_1),\psi(c_2)) = ((1+3^\alpha)/2)^{\frac{1}{\alpha}}$ and $\mu_\alpha(\psi(c_3),\psi(c_4))=2$. Due to the convexity of the power function when $\alpha>1$, then $\mu_\alpha(\psi(c_1),\psi(c_2)) > \mu_\alpha(\psi(c_3),\psi(c_4))$ and consequently $\widetilde{\sigma}_\alpha(c_1,c_2) < \widetilde{\sigma}_\alpha(c_3,c_4)$. When $\alpha<1$, the inequality is inverted.

Proposition 3. The similarities of the family $\{\widetilde{\sigma}_\alpha\}_{\alpha\in\mathbb{R}}$ are decreasing functions of α.

This is due to the fact that the α-means are increasing functions of α (e.g., [5]).

5 Experimental results

In this section, we present two complementary comparisons based on the subsumption hierarchy of WordNet 2.0 [8]. First, we compare the information content restricted to the structural information with the well-known Resnik's information content which additionally requires a corpus. This allows us to quantify the information deduced from the corpus. Second, we use three well-known benchmarks (Rubenstein & Goodenough [22], Miller & Charles [16], Finkelstein et al. [9]) which gather human judgments on some concept pairs. This allowed us to evaluate the relevance of the different approximations.

5.1 Comparison on WordNet

This subsection presents a comparison between the information content based on different approximations. We restrict ourselves to nouns and to the subsumption hierarchy (hyperonymy/hyponymy) of WordNet. This hierarchy which contains 146690 nodes constitutes the backbone of the noun subnetwork accounting for close to 80% of the links [3]. The computations have been performed with the Perl modules of Pedersen et al. [17] which allowed us to adapt tree-based measures to the WordNet structure. Hence, although a synset could have more than one hyperonym, we have represented it as a tree model $T_{WordNet}(\mathcal{C})$. We have also added some Perl modules to take into account all the new approximations presented in this paper. The main interest of $T_{WordNet}(\mathcal{C})$ is to be large enough to allow computations of robust statistics and we do not enter here into the discussion between experts concerning the ontological nature of WordNet.

We have computed the information content for four different concept sets: the whole set of WordNet (146690 concepts) and three subsets of WordNet composed of the concept sets used respectively in the Miller & Charles [16], Rubenstein & Goodenough [22] and Finkelstein & Gabrilovich [9] benchmarks. We have compared the approximations $\widehat{\mathrm{P}}_d$, $\widehat{\mathrm{P}}_g$ and $\widehat{\mathrm{P}}_r$. The correlations $\rho(\psi_d,\psi_r)$ and $\rho(\psi_g,\psi_r)$ are reported in the figure 1 (the rank correlations not reported here give similar results).

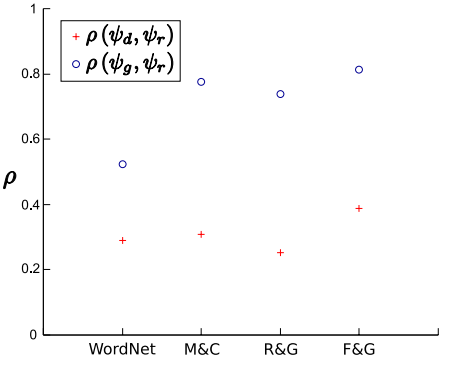

Figure 1. Correlation of ψ_d and ψ_g information content with the one of Resnik ψ_r on WordNet concepts and four subsets

The approximation $\widehat{\mathrm{P}}_r$ which is a yardstick has been computed with the British National Corpus with the Resnik counting method and a smoothing by 1 [17].

We can remark that each benchmark uses a sample of concepts which is not so representative of the whole set of concepts. Indeed, the corpus effect on the information content is more important on the whole set than on the three samples. From this point of view, the one of Finkelstein & Gabrilovich is the worse benchmark.

Unsurprisingly, the information content based on the approximation $\widehat{\mathrm{P}}_d$ is the less correlated with $\widehat{\mathrm{P}}_r$. However, the positive correlations show the relationship between the ascending and descending approximations: the depth tends to be conversely proportional to the height.

The correlations between ψ_g and ψ_r show that the information quantity deduced from the corpus is restricted comparatively to the information deduced from the hierarchical structure. Nevertheless, these results depend on the corpus and the structure of WordNet. That's why further work is required to generalize this conclusion to a large set of ontologies.

5.2 Comparisons with human judgments

As showed in section 3.1, two components are essential when comparing two concepts c_i and c_j: the shared information content ($\psi^\cap(c_i,c_j) = \psi(c_{ij})$) and the distinguishing information content ($\psi^\triangle(c_i,c_j) = \psi(c_i)+\psi(c_j)-2\cdot\psi(c_{ij})$). To measure the specific influence of these two components we have computed the correlation of each of them with the human judgment. The considered human judgment evaluations are taken from the Miller & Charles [16], Rubenstein & Goodenough [22], Finkelstein & Gabrilovich [9] experiments and the approximation of P is the Resnik's approximation. The results (figure 2) closely depend on the test sets.

The contribution of $\psi_r^\triangle$ is more important than the one of $\psi_r^\cap$ for the benchmarks of Miller & Charles and Rubenstein & Goodenough

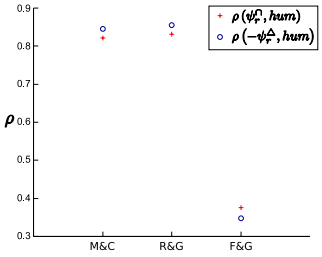

Figure 2. Contribution of $\psi^{\cap}$ and $\psi^{\triangle}$ with $\widehat{\mathrm{P}}_r$ to simulate human judgment

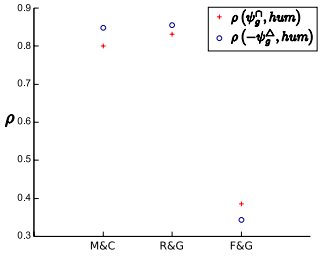

Figure 3. Contribution of $\psi^{\cap}$ and $\psi^{\triangle}$ with $\widehat{\mathrm{P}}_g$ to simulate human judgment

contrary to the Finkelstein & Gabrilovich benchmark. This tend to express the variability of human sensibility which can be due to the evaluation process of the three benchmarks.

Moreover the previous experiments have shown that $\widehat{\mathrm{P}}_g$ seems to be the more efficient (better correlated with human judgments) approximation comparing to the Resnik's approximation which uses a corpus. Hence, we have computed the correlations of the two components $\psi^{\cap}$ and $\psi^{\triangle}$ with the human judgment with $\widehat{\mathrm{P}}_g$ (figure 3). The results are very similar to those obtained with the Resnik's approximation. This tend to suppose that the information deduced from the corpus contain as much information as noise.

6 Conclusion

The concept of similarity is fundamental in numerous fields (e.g., classification, AI, psychology, ...). At the origin, the definitions are often built to fulfill precise objectives in specific domains. However, several measures (e.g., [12, 7]) have shown their relevance to very different applications. Nowdays similarities know a significant renewed interest associated with the expansion of the ontologies in knowledge engineering. In this framework, the most often used measures to quantify proximities between concept pairs are tree-based similarities whose definitions may integrate or not additional information from a textual corpus. In practice, the choice of a similarity is a critical step since the results of the algorithms often closely depend on this choice.

In this paper, we have built a new theoretical framework which allows to rewrite homogeneously numerous similarity functions used in knowledge engineering. We believe that such an approach, in the spirit of the pioneer work of Lin, is important for two major reasons. First, this rewriting highlights relationships both semantically and structurally between a large set of measures which have been originally defined for very different purposes. And, it has allowed to deduce mathematical properties. Second, it can guide the proposition of new measures by making explicit the information on the ontology which has not been integrated into the definitions yet. In this way, we have here proposed new approximations which allow to better exploit the information associated with the hierarchical structure of the ontology.

We have also restricted ourselves to similarities for subsumption hierarchies without multiple inheritance. We have started to extend our approach to subsumption hierarchy with multiple inheritance.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper.

REFERENCES

[1] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. Owl web ontology language reference, 2004. http://www.w3.org/TR/owl-ref/.
[2] E. Blanchard, P. Kuntz, M. Harzallah, and H. Briand, 'A tree-based similarity for evaluating concept proximities in an ontology', in *Proc. 10th Conf. Int. Federation Classification Soc.*, pp. 3–11. Springer, (2006).
[3] A. Budanitsky, 'Lexical semantic relatedness and its application in natural language processing', Technical report, Univ. of Toronto, (1999).
[4] A. Budanitsky and G. Hirst, 'Evaluating wordnet-based measures of semantic distance', *Computational Linguistics*, **32**(1), 13–47, (2006).
[5] P.S. Bullen, D. S. Mitrinovic, and P. M. Vasics, *Means and their inequalities*, Reidel, 1988.
[6] F. Caillez and P. Kuntz, 'A contribution to the study of the metric and euclidean structures of dissimilarities', *Psychometrika*, **61**(2), 241–253, (1996).
[7] L. R. Dice, 'Measures of the amount of ecologic association between species', *Ecology*, **26**(3), 297–302, (1945).
[8] *WordNet: An electronic lexical database*, ed., C. Fellbaum, MIT Press, 1998.
[9] L. Finkelstein, E. Gabrilovich, Y. Matias, G. Wolfman E. Rivlin, Z. Solan, and E. Ruppin, 'Placing search in context: The concept revisited', *ACM Trans. Information Systems*, **20**(1), 116–131, (2002).
[10] J.C. Gower and P. Legendre, 'Metric and euclidean properties of dissimilarity coefficients', *J. of Classification*, **3**, 5–48, (1986).
[11] Z. Hubalek, 'Coefficient of association and similarity based on binary (presence, absence) data: an evaluation', *Biological Reviews*, **57**(4), 669–689, (1982).
[12] P. Jaccard, 'Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines', *Bulletin de la Société Vaudoise de Sciences Naturelles*, (37), 241–272, (1901). (in french).
[13] J. H. Lee, M. H. Kim, and Y. J. Lee, 'Information retrieval based on conceptual distance in is-a hierarchies', *J. Documentation*, **49**(2), 188–207, (1993).
[14] D. Lin, 'An information-theoretic definition of similarity', in *Proc. 15th Int. Conf. Machine Learning*, pp. 296–304. Morgan Kaufmann, (1998).
[15] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani, 'Algorithmic detection of semantic similarity', in *Proc. 14th Int. Conf. World Wide Web*, pp. 107–116. ACM Press, (2005).
[16] G.A. Miller and W.G. Charles, 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, **6**(1), 1–28, (1991).
[17] T. Pedersen, S. Patwardhan, and J. Michelizzi, 'Wordnet similarity - measuring the relatedness of concepts', in *Proc. 5th Ann. Meet. North American Chapter Assoc. Comp. Linguistics*, pp. 38–41, (2004).
[18] R. Rada, H. Mili, E. Bicknell, and M. Blettner, 'Development and application of a metric on semantic nets', *IEEE Trans. Syst., Man, Cybern.*, **19**(1), 17–30, (1989).
[19] P. Resnik, *Selection and Information : A Class based Approach to Lexical Relationships*, Ph.D. dissertation, University of Pennsylvania, 1993.
[20] P. Resnik, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *J. Artificial Intell. Research*, **11**, 95–130, (1999).
[21] E. Rosch, 'Cognitive representations of semantic categories', *Experimental Psychology: Human Perception and Performance*, **1**, 303–322, (1975).
[22] H. Rubenstein and J.B. Goodenough, 'Contextual correlates of synonymy', *Comm. ACM*, **8**(10), 627–633, (1965).
[23] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, 'Content-based image retrieval at the end of the early years', *IEEE Trans. Pattern Anal. Machine Intell.*, **22**(12), 1349–1380, (2000).
[24] R. R. Sokal and P. H. Sneath, *Principles of numerical taxonomy*, W. H. Freeman, 1963.
[25] O. Steichen, C. Daniel-Le Bozec, M. Thieu, E. Zapletal, and M.-C. Jaulent, 'Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus', *Computers in Biology and Medicine*, **36**(7-8), 768–788, (2006).
[26] N. Stojanovic, A. Maedche, S. Staab, R. Studer, and Y. Sure, 'Seal: a framework for developing semantic portals', in *Proc. Int. Conf. Knowledge Capture*, pp. 155–162, (2001).
[27] A. Tversky, 'Features of similarity', *Psychological Review*, **84**(4), 327–352, (1977).
[28] Z. Wu and M. Palmer, 'Verb semantics and lexical selection', in *Proc. 32nd Annual Meeting Assoc. Computational Linguistics*, pp. 133–138, (1994).