A DISTANCE FUNCTION TO ASSESS THE SIMILARITY OF WORDS USING ONTOLOGIES

Montserrat Batet[1] **Aida Valls**[1] **Karina Gibert**[2]

[1] Universitat Rovira i Virgili, Department of Computer Science and Mathematics
Av. Països Catalans 26, 43007 Tarragona, Spain
{montserrat.batet,aida.valls}@urv.cat
[2] Universitat Politècnica de Catalunya, Department of Statistics and Operations Research
Campus Nord, Ed.C5, c/Jordi Girona 1-3, 08034 Barcelona, Spain
karina.gibert@upc.edu

Abstract

When comparing categorical values, traditional approaches use metrics based only on the matching of the values, obtaining a Boolean result. In this paper, it is proposed to use a measure able to compute the degree of semantic similarity between a pair of terms using an ontology as background knowledge. The presented measure - the Superconcept-based distance - have two main advantages over other approaches based on the exploitation of the hierarchical model of ontologies: on one hand, it takes into account the whole hierarchy of concepts in the ontology to assess the similarity between a pair of words; on the other hand, this paper proves that this measure fulfills the distance properties. As the paper also reviews, the usual semantic similarity measures used in the literature does not fulfill the triangle inequality, which prevents them from being used in some decision making methods.

Keywords: Semantic similarity, ontologies, linguistic terms, distance properties.

1 INTRODUCTION

Our work is focused on decision making problems that have to deal with variables that take their values in a list of linguistic terms. Differently from categorical variables that have a predefined domain of terms (i.e. modalities), we are facing the case of having a non fixed and large set of possible values. Moreover, no ordering or measurement scale in the values is defined, as traditional linguistic variables do. An example of this type of variables can be languages or hobbies.

Traditionally, the comparison between two values in categorical variables is done simply based on the equality/inequality of the words, due to the lack of proper methods for representing the meaning of the terms. Some widely used distance measures for categorical values are the Chi-Squared and the Hamming distance [16]. However, from a semantic point of view, it is possible to establish different degrees of similarity between values (i.e. Italian is more similar to French than to Chinese). Each of these terms is in fact describing a concept, thus, reasoning at a conceptual level should be done in order to calculate an approximation of the similarity.

The computation of the semantic similarity between concepts is an active trend in computational linguistics. The similarity between a pair of concepts quantifies how they are alike based on the estimation of semantic evidence observed in some knowledge source. Taxonomies and, more generally ontologies [13] are considered as a graph model in which semantic relations are modeled as links between concepts. In the literature several semantic similarity measures have been proposed. A brief review of those measures will be presented in section 2. In this paper we focus on measures based on the exploitation of the taxonomic relations in ontologies. These measures are based on the computation of the minimum path length between a pair of concepts to assess the semantic similarity. Consequently, a lot of relevant information is missed because the rest of taxonomic information is not considered.

To sort out this problem, in this paper a new measure is presented, which takes into account the common and not common ancestors (i.e. superconcepts) of the two concepts compared. It is called Superconcept-based Distance (SCD). When this comparison between terms is done in the context of decision making, it is worth to know the metrical properties of the measure, because it may have implications on the results that will be obtained. For example, for the particular case of hierarchical clustering, it is interesting to maintain the ultrametric properties of the dendograms and the Huygens theorem of decomposition of inertia, which are directly related with the interpretability of the final results. If the comparison measure is a distance, these properties are guaranteed [3]. However, in the field of computational linguistics, the comparison measures proposed usually do not fulfill the triangle inequality property, being only similarities but not distances. In this field, this is not a handicap. As noted by Tversky [14], the triangle inequality property is not always achievable when dealing with linguistic descriptions.

The goal of this paper is to study the Superconcept-based measure and analyze its metrical properties, proving that it is a distance. Since most of the classical semantic similarity measures are not distances, this is an interesting result that may help to extend the use of this similarity assessment not only to text mining but also to other fields, specially to intelligent decision support systems.

The paper is organized as follows. Section 2 reviews the existing semantic similarity measures based on ontologies. Section 3 defines the Superconcept-based Distance. Section 4 shows the good performance results of this measure. Section 5 proves the distance properties of SCD. Discussion and future work are in section 6.

2 RELATED WORK

In the literature, we can found several different approaches to compute semantic similarity between concepts according to the techniques employed and the knowledge exploited. There are approaches in which the degree of co-occurrence between terms in a given corpus is used as an estimation of similarity [7]. These measures need a corpus as general as possible in order to estimate social-scale word usage. However, due to the lack of domain coverage of a general domain corpus and the difficulty of compiling a relevant domain corpus big enough to obtain robust statistics, they offer a limited performance.

Similarity computation can be also based on structured representations of knowledge, typically, subsumption hierarchies. The evolution of those hierarchies have given the origin to ontologies in which many types of relationships and logical descriptions can be specified to formalize knowledge [9].

Some approaches combine the knowledge provided by an ontology with the Information Content (IC) of the concepts that are being compared. IC measures the amount of information provided by a given term from its probability of appearance in a corpus. Based on this premise, Resnik [11] proposed to estimate the similarity between two terms as the amount of information they share in common. In a taxonomy, this information is represented by the Least Common Subsumer (LCS) of both terms. So, the computation of the IC of the LCS results is used to estimate the similarity of the subsumed terms. The more specific the subsumer is (higher IC), the more similar the subsumed terms are. Several variations of this measure have been developed [5]. However, those measures are affected by the availability of the background corpus and their coverage with respect to the evaluated terms.

Other approaches consider ontologies as a graph model in which semantic interrelations are modeled as links between concepts. These measures compute concept similarity as inter-link distance in the taxonomic structure (also called Path Length) [10].

$$sim_{pL}(c_1,c_2)=min\ \#\ of\ is-a\ edges\ connecting\ c_1\ and\ c_2 \quad (1)$$

Several variations of this measure have been developed such as the one proposed by Wu and Palmer [15]. Their proposal also takes into account the depth of the concepts in the hierarchy (2). Here, N_1 and N_2 is the number of is-a links from c_1 and c_2 respectively to the LCS c, and N_3 is the number of is-a links from c to the root ρ of the ontology.

$$sim_{w\&p}(c_1, c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (2)$$

Leacock and Chodorow [6] proposed a measure that considers both the number of nodes N_p from c_1 to c_2, and the depth D of the taxonomy in which they occur (3).

$$sim_{l\&c}(c_1, c_2) = -\log N_p/2D \quad (3)$$

However, those measures based on Path Length only consider the minimum path between a pair of concepts, omitting the rest of the taxonomic knowledge available in the ontology, wasting a great amount of relevant knowledge.

3 SUPERCONCEPT BASED DISTANCE

The Superconcept-based Distance (SBC) has been designed to overcome the limitations of the other Path Length approaches, explained in the previous section. This measure takes into account both the amount of overlapping and non-overlapping taxonomical knowledge between concepts. To do this, the distance considers the number of shared superconcepts (upper concepts of a concept in the taxonomy) and the number of non shared superconcepts.

Considering that an ontology O as an object model composed by a set of concepts or classes C, which are *taxonomically* related by the transitive is-a relation $H^c \in C \times C$, called concept hierarchy or taxonomy, and *non-taxonomically* related by named object relations $R^* \in C \times C\times$ String, the Superconcept-based Distance is defined as:

Definition: Superconcept-based Distance (SCD)

$$d_{\mathcal{SCD}}(c_i, c_j) = \sqrt{\frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|}} \qquad (4)$$

where $\mathcal{A}(c_i) = \{c_j \in C | c_j$ is superconcept of $c_i \in H^C \vee \ c_i = c_j\}$.

The numerator of expression (4) calculates the number of non-common ancestors of the two concepts, c_i and c_j. That is, the amount of non-shared information. The denominator is used to consider the proportion of non-common superconcepts with respect to the total number of superconcepts. In this way, the measure is able to distinguish between cases in which the number of common superconcepts between a pair of concepts is small from those cases in which the number of common superconcepts is high. Less shared superconcepts means a greater distance. The squared root is used to smooth the result as the amount of common information is not linear in relation to similarity [7].

4 EVALUATION

The most common way of evaluating similarity measures is by using a set of word pairs whose similarity has been assessed by a group of human experts and calculate the correlation between the human ratings and the results of the computerized measures. The most commonly used benchmarks are the one proposed by Miller and Charles [8], which is a set of 30 domain-independent word pairs chosen from their hight, middle and low level of synonymy from an experiment done by Rubenstein and Goodenough [12], and the one proposed by Resnik [11]. Resnik replicated the experiment of Miller and Charles in order to obtain more accurate ratings by giving a group of experts the same set of noun pairs. Resnik computed how well the rating of the subjects in his evaluation correlated with Miller and Charles ratings. The average correlation over the 10 subjects (considered by Resnik) was 0.884. This value is considered the upper bound to what one could expect from a machine computation on the same task [11]. In order to evaluate the measure presented, we used these benchmarks and WordNet[4] as background ontology (Table 1).

The results show that Leacock & Chodorow and Wu & Palmer approaches clearly outperform Path Length measure. The reason is that these measures take into account more information of the taxonomy as the depth of the ontology (Leacock and Chodorow's proposal) or the the depth of the LCS and the relative depth between compared concepts and the LCS (Wu and Palmer). Our measure (SCD) considers the whole set of superconcepts of the pair of compared concepts. It can be seen that our proposal provides the highest correlation regarding expert's judgments. Notice that for the Resnik benchmark, with SCD measure we are approaching the upper bound (0.884) to what one could expect from a machine computation.

Table 1: Correlations obtained for each measure using domain independent benchmarks

Measure	Resnik	M&C	Rubenstain
Path Length	0.670	0.638	0.737
Wu&Palmer	0.804	0.795	0.801
Leacock&Chodorow	0.829	0.779	0.838
SCD	**0.839**	**0.807**	**0.839**

The distance had been also evaluated for the biomedical domain using a benchmark referring to medical disorders and SNOMED-CT terminology [1]. In this study, SCD also obtained the best performance results.

5 PROOF OF THE DISTANCE PROPERTIES

The properties that a distance measure must fulfill are the following ones:

1. Identity: $d(c_i, c_j) = 0 \Longleftrightarrow c_i = c_j, \forall i, j$
2. Symmetry: $d(c_i, c_j) = d(c_j, c_i), \ \forall i, j$
3. Triangle inequality: $d(c_i, c_j) + d(c_j, c_k) \geq d(c_i, c_k), \ \forall i, j, k$

The proofs of the Identity and Symmetry are straightforward. The Identity is fulfilled because iff two concepts are the same, the union of all the ancestors is equal to the set of common ancestors (i.e. intersection), which leads to a 0 at the numerator of eq. 4. Symmetry also holds because the union and intersection operators are symmetric.

So, in this section we concentrate on the proof of the the triangle inequality property. First some notation is introduced to simplify the formulation of the problem.

Let:

$$A = \mathcal{A}(c_1), B = \mathcal{A}(c_2), C = \mathcal{A}(c_3)$$

Then, the triangle inequality property is expressed as:

$$\overline{\frac{|A \cup B| - |A \cap B|}{|A \cup B|}} + \overline{\frac{|B \cup C| - |B \cap C|}{|B \cup C|}} \geq \overline{\frac{|A \cup C| - |A \cap C|}{|A \cup C|}} \qquad (5)$$

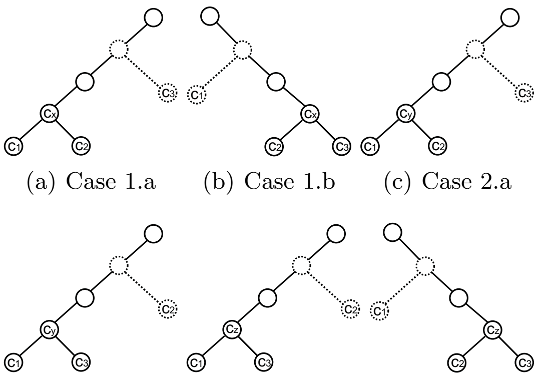
Proposition 1. In a taxonomic relation without multiple inheritance, this property holds:

$A \cap B = B \cap C$ or $A \cap B = A \cap C$ or $B \cap C = A \cap C$

Proof. Assuming that proposition 1 is not true, the following conditions must be true simultaneously:

$$\begin{array}{rl} 1.\ A \cap B \neq B \cap C \Rightarrow & (1.a)\ \exists c_x \in A \cap B, c_x \notin B \cap C \\ \text{or} & (1.b)\ \exists c_x \notin A \cap B, c_x \in B \cap C \\ 2.\ A \cap B \neq A \cap C \Rightarrow & (2.a)\ \exists c_y \in A \cap B, c_y \notin A \cap C \\ \text{or} & (2.b)\ \exists c_y \notin A \cap B, c_y \in A \cap C \\ 3.\ B \cap C \neq A \cap C \Rightarrow & (3.a)\ \exists c_z \in A \cap C, c_z \notin B \cap C \\ \text{or} & (3.b)\ \exists c_z \notin A \cap C, c_z \in B \cap C \end{array}$$

These six possible cases are graphically represented in figure 1.



(a) Case 1.a (b) Case 1.b (c) Case 2.a

(d) Case 2.b (e) Case 3.a (f) Case 3.b

Figure 1: Intersection cases

From these figures, it is easy to see that:

$$\begin{array}{c} \text{If expr. } 1.a = \mathit{true} \Rightarrow B \cap C = A \cap C \text{ (Fig. 1(a))} \\ \text{If expr. } 1.b = \mathit{true} \Rightarrow A \cap B = A \cap C \text{ (Fig. 1(b))} \\ \text{If expr. } 2.a = \mathit{true} \Rightarrow B \cap C = A \cap C \text{ (Fig. 1(c))} \\ \text{If expr. } 2.b = \mathit{true} \Rightarrow A \cap B = B \cap C \text{ (Fig. 1(d))} \\ \text{If expr. } 3.a = \mathit{true} \Rightarrow A \cap B = B \cap C \text{ (Fig. 1(e))} \\ \text{If expr. } 3.b = \mathit{true} \Rightarrow A \cap B = A \cap C \text{ (Fig. 1(f))} \end{array}$$

□

Proof. Triangle inequality.

For simplicity, let us rename the sets of union and intersection as follows:

$$\begin{array}{c} x = |A \cup B| \text{ and } p = |A \cap B| \text{ with } 1 \leq p \leq x \\ y = |B \cup C| \text{ and } q = |B \cap C| \text{ with } 1 \leq q \leq y \\ z = |A \cup C| \text{ and } r = |A \cap C| \text{ with } 1 \leq r \leq z \end{array}$$

Notice that $p = 1$ iff $A \cap B$ =root node of the ontology, $q = 1$ iff $B \cap C$ =root node of the ontology and $r = 1$ iff $A \cap C$ =root node of the ontology. In addition, $p = x$ iff $c_1 = c_2$, $q = x$ iff $c_2 = c_3$ and $r = x$ iff $c_1 = c_3$.

So the triangle inequality can be expressed as:

$$\sqrt{\frac{x-p}{x}} + \sqrt{\frac{y-q}{y}} \geq \sqrt{\frac{z-r}{z}}$$

, that is:

$$\frac{\sqrt{(x-p)yz}}{\sqrt{xyz}} + \frac{\sqrt{(y-q)xz}}{\sqrt{xyz}} \geq \frac{\sqrt{(z-r)xy}}{\sqrt{xyz}}$$

This is equivalent to demonstrate:

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyr} \qquad (6)$$

Let us prove eq.6 in the three cases of Proposition 1.

1. If $A \cap B = B \cap C \Rightarrow p = q$ and $p \leq r$

$$\sqrt{xyz - pyz} + \sqrt{xyz - xpz} \geq \sqrt{xyz - xyr}$$

$$(\sqrt{xyz - pyz} + \sqrt{xyz - xpz})^2 \geq (\sqrt{xyz - xyr})^2$$

$$xyz + xyr - pz(x+y) + 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq 0$$

In this case we can rewrite x, y and z as:

$$\begin{array}{rcl} x & = & a + b + p \\ y & = & b + c + p \\ z & = & a - (r - p) + c - (r - p) + r \\ & = & a + 2p + c - r \end{array}$$

,where $a = |A| - p$, $b = |B| - p$, and $c = |C| - p$. So, a, b, c are the number of superconcepts of A, B, C, respectively, that not belong to $A \cap B$. Then,

$$\begin{array}{rl} & a^2b + 4abp + 2abc - abr + a^2c + 4acp + ac^2 \\ - & acr + a^2p + 3ap^2 - apr + b^2a + 2b^2p + b^2c \\ - & b^2r + 4bcp + bc^2 - bcr + 4bp^2 - 2bpr + 3cp^2 \\ + & c^2p - cpr + 2p^3 - p^2r + abr + acr + apr \\ + & b^2r + bcr + 2bpr + cpr + p^2r - a^2p - 2abp \\ - & 4ap^2 - 2acp - 4bp^2 - 4p^3 - 4cp^2 - 2bcp \\ - & c^2p + apr + 2bpr + 2p^2r + cpr \\ + & 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq 0 \end{array}$$

Having that $p \leq r$ (the number of concepts of $A \cap B$ is less or equal than the number of concepts of $A \cap C$) and $p \leq x$ (the number of concepts of $A \cap B$ is less or equal than the number of concepts of $A \cup B$) and $p \leq y$ (the number of concepts of

$B \cap C$ is less than number of concepts of $B \cup C$), we can determine the positivity of the pairs:

$$\begin{aligned}
& a^2b + 2abp + 2abc + a^2c + 2acp + ac^2 \\
+ \quad & b^2a + 2b^2p + b^2c + 2bcp + bc^2 \\
+ \quad & \underbrace{2p^2r - 2p^3}_{\geq 0} + \underbrace{apr - ap^2}_{\geq 0} + \underbrace{cpr - cp^2}_{\geq 0} + 2bpr \\
+ \quad & 2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xpz}_{\geq 0}} \geq 0
\end{aligned}$$

So, triangle inequality holds when $A \cap B = B \cap C$.

2. If $A \cap B = A \cap C \Rightarrow p = r$ and $p \leq q$

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyp}$$

$$xyz + yp(x - z) - xqz + 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

In this case we can rewrite x, y and z as:

$$\begin{aligned}
x &= a + b + p \\
y &= b - (q - p) + c - (q - p) + q \\
&= b + 2p + c - q \\
z &= a + c + p
\end{aligned}$$

,where $a = |A| - p$, $b = |B| - p$, and $c = |C| - p$. Then,

$$\begin{aligned}
& a^2b + 2a^2p + a^2c - a^2q + 2abc + 4acp \\
+ \quad & ac^2 - acq + 4abp + 4ap^2 - 2apq + ab^2 \\
- \quad & abq + b^2c + 4bcp + bc^2 - bcq + b^2p + 3bp^2 \\
- \quad & bpq + 3cp^2 + c^2p - cpq + 2p^3 - p^2q + b^2p \\
+ \quad & 2bp^2 - 2cp^2 - c^2p - bpq + cpq - a^2q - acq \\
- \quad & 2apq - abq - bcq - bpq - cpq - p^2q \\
+ \quad & 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0
\end{aligned}$$

In this case, the number of superconcepts of c_2 and c_3 are greater or equal than $|B \cap C|$, so we have that $q \leq c + p$ and $q \leq b + p$. In addition, we have that $p \leq x$ and $q \leq y$. So:

$$\begin{aligned}
& \underbrace{a^2p + a^2b - a^2q}_{\geq 0} + \underbrace{a^2p + a^2c - a^2q}_{\geq 0} \\
+ \quad & \underbrace{acp + ac^2 - acq}_{\geq 0} + \underbrace{acp + abc - acq}_{\geq 0} \\
+ \quad & \underbrace{2ap^2 + 2acp - 2apq}_{\geq 0} + \underbrace{2ap^2 + 2abp - 2apq}_{\geq 0} \\
+ \quad & \underbrace{abp + ab^2 - abq}_{\geq 0} + \underbrace{abp + abc - abq}_{\geq 0} \\
+ \quad & \underbrace{bcp + b^2c - bcq}_{\geq 0} + \underbrace{bcp + bc^2 - bcq}_{\geq 0} \\
+ \quad & \underbrace{2bp^2 + 2b^2p - 2bpq}_{\geq 0} + \underbrace{bp^2 + bcp - bpq}_{\geq 0} \\
+ \quad & \underbrace{2p^3 + 2bp^2 - 2p^2q}_{\geq 0} + \underbrace{cp^2 + bcp^2 - cpq}_{\geq 0} \\
+ \quad & 2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xqz}_{\geq 0}} \geq 0
\end{aligned}$$

So, the triangle inequality is demonstrated when $A \cap B = A \cap C$.

3. If $B \cap C = A \cap C \Rightarrow q = r$ and $q \leq p$

$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyq}$$

$$xyz + xq(y - z) - pyz + 2\sqrt{xyz - pyz}\sqrt{xyz - xqz} \geq 0$$

In this case we can rewrite x, y and z as:

$$\begin{aligned}
x &= a - (p - q) + b - (p - q) + p = \\
&= a + b + 2q - p \\
y &= b + c + q \\
z &= a + c + q
\end{aligned}$$

,where $a = |A| - p$, $b = |B| - p$, and $c = |C| - p$ and $q \leq p$. And having that $p \leq a + q$, $p \leq b + q$, $p \leq x$ and $q \leq y$:

$$\begin{aligned}
& \underbrace{a^2b + abq - abp}_{\geq 0} + \underbrace{ab^2 + abq - abp}_{\geq 0} \\
+ \quad & \underbrace{abc + bcq - bcq}_{\geq 0} + \underbrace{b^2c + bcq - bcp}_{\geq 0} \\
+ \quad & \underbrace{2b^2q + 2bq^2 - 2bqp}_{\geq 0} + \underbrace{abq + bq^2 - bqp}_{\geq 0} \\
+ \quad & \underbrace{a^2c + acq - acp}_{\geq 0} + \underbrace{abc + acq - acp}_{\geq 0} \\
+ \quad & \underbrace{ac^2 + c^2q - c^2p}_{\geq 0} + \underbrace{bc^2 + c^2q - c^2p}_{\geq 0} \\
+ \quad & \underbrace{2acq + 2cq^2 - 2cqp}_{\geq 0} + \underbrace{2bcq^2 + 2cq^2 - 2cqp}_{\geq 0} \\
+ \quad & \underbrace{abq + aq^2 - aqp}_{\geq 0} + \underbrace{2bq^2 + 2q^3 - 2q^2p}_{\geq 0} \\
+ \quad & 2\sqrt{\underbrace{xyz - pyz}_{\geq 0}}\sqrt{\underbrace{xyz - xqz}_{\geq 0}} \geq 0
\end{aligned}$$

Finally, the triangle inequality is demonstrated when $B \cap C = A \cap C$. □

6 DISCUSSION AND FUTURE WORK

This paper proposes the use of semantic-based similarity measures for the comparison on linguistic terms in decision making. In particular, the measures based on the exploitation of the taxonomic relations in an ontology have been reviewed. The main drawbacks of these measures are that they only exploit the minimum path length between a pair of concepts of an ontology, wasting the rest of taxonomic relations, and that they do not hold the triangle inequality property.

In this paper, the SCD measure that takes into account the ratio between the shared and non-shared taxonomically related concepts of the compared pair of concepts is analyzed. As shown in the evaluation, our proposal outperforms previous attempts, exploiting the taxonomical structure of WordNet.

The paper proves that SCD measure fulfils the metric properties, which means that is suitable to be used in some applications, such as in hierarchical clustering.

In fact, an approximation using semantic measures based on ontologies to make clustering was studied in [2]. In that work, it was demonstrated that the clusters obtained using these measures are more accurate and interpretable.

As future work, we plan to evaluate the SCD measure with some domain ontologies and using other domain dependent benchmarks, which it will be interesting in order to evaluate the dependency of the results in relation to the ontology coverage. We also plan to extend the use of this semantic-based comparison of terms to other decision making methods.

Acknowledgements

This work has been partially funded by the Spanish research projects DAMASK (TIN2009-11005) and ARES (CSD2007-00004) and the Universitat Rovira i Virgili (2009 AIRE-04). Montserrat Batet is also supported by a research grant provided by the Universitat Rovira i Virgili. The authors want to specially thank the help of Dr. Juan A. Rodriguez.

References

[1] M. Batet, D. Snchez, A. Valls, and K. Gibert. Ontology-based semantic similarity in the biomedical domain. In *12th Conference on Artificial Intelligence in Medicine (AIME'09). WS. Intelligent Data Analysis in Biomedicine and Pharmacology*, Italy, 2009.

[2] M. Batet, A. Valls, and K. Gibert. Improving classical clustering with ontologies. In *Proceedings of the 4th World conference of the IASC*, number 137–146, Japan, 2008.

[3] W. R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* Wiley, 1984.

[4] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press. More information: http://www.cogsci.princeton.edu/ wn/, Cambridge, Massachusetts, 1998.

[5] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the International Conference on Research in Computational Linguistics*, pages 19–33, Sep 1997.

[6] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, chapter WordNet: An electronic lexical database, pages 265–283. MIT Press, 1998.

[7] B. Lemaire and G. Denhire. Effects of high-order co-occurrences on word semantic similarities. *Current Psychology Letters - Behaviour, Brain and Cognition*, 18(1), 2006.

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

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

[10] R. Rada, H. Mili, E. Bichnell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):17–30, 1989.

[11] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, pages 448–453, Montreal, Canada, 1995.

[12] H. Rubenstein and J. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[13] R. Studer, V. R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *IEEE Transactions on Data and Knowledge Engineering*, 25(1-2)(1-2):161–197, 1998.

[14] A. Tversky. Features of similarity. *Phychological Review*, 84:327–352, 1977.

[15] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proc. of the 32nd annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico, USA, 1994.

[16] R. Xu and D. Wunsch II. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, May 2005.