Theoretical Formulas of Semantic Measure: A Survey

Kalthoum Rezgui
Higher Institute of Management/SOIE Laboratory, Tunis, Tunisia
Email: kalthoum.rezgui@isg.rnu.tn

Hédia Mhiri and Khaled Ghédira
Higher Institute of Management/SOIE Laboratory, Tunis, Tunisia
Email: {hedia.mhiri, khaled.ghedira}@ isg.rnu.tn

***Abstract*—In recent years, several semantic similarity and relatedness measures have been developed and applied in many domains including linguistics, biomedical informatics, GeoInformatics, and Semantic Web. This paper discusses different semantic measures which compute similarity and relatedness scores between concepts based on a knowledge representation model offered by ontologies and semantic networks. The benchmarks and approaches used for the evaluation of semantic similarity methods are also described. The aim of this paper is to give a comprehensive view of these measures which helps researchers to choose the best semantic similarity or relatedness metric for their needs.**

***Index Terms*— semantic similarity, semantic relatedness, ontology, semantic Web, WordNet**

I. Introduction

In the literature, several works on semantic measures have been proposed to compute similarity or relatedness scores between concepts based on semantic networks and ontologies. Some of them explore path lengths among nodes in the hierarchy, others consider in addition the position or depth of nodes in the hierarchy, others rely on statistical analysis of corpora to associate probabilities with concepts in order to compute information content represented by nodes while a last group exploits textual descriptions of concepts in dictionaries. In this paper, we start by presenting a classification of semantic measures. Then, in Section 3 and 4, we present and discuss the different approaches related to the problem of computation of a semantic (similarity/relatedness) score in a knowledge representation model, particularly the case of knowledge modeled in the form of a concept hierarchy. In Section 5, we describe well-known benchmarks and broad evaluation approaches that are mostly used for assessing the quality of semantic measures. Finally, conclusion is presented in Section 6.

II. Concept of Similarity, Distance or Semantic Relatedness

In the literature, three main classes of semantic measures between concepts are commonly quoted:

- **Semantic similarity** when the measure computes whether two concepts are semantically similar, that is, they share common properties and attributes.
- **Semantic relatedness** when the measure computes whether two concepts are semantically related, that is, they are connected in their function. It is considered as a general case of semantic similarity in the works of Resnik [1] and Budanitsky and Hirst [2].
- **Semantic distance** when the measure computes whether two concepts are semantically distant. According to [2], semantic distance is the inverse of semantic relatedness. The idea behind this is that "*the more two terms are semantically related, the more semantically close they are*" [2].

III. Semantic Similarity Measures

In general, the works dealing with semantic similarity measures can be classified into three families of approaches: edge-based approaches, node-based approaches or information-theoretic approaches, and hybrid approaches. Most of these methods exploit particular lexical resources, such as dictionaries, corpus, or well structured taxonomies.

A. Edge-based Approaches

This category of semantic measure approaches is based on the length of paths in a tree to determine the distance between two given concepts. In what follows, we present the similarity measures of Rada et al. [3], Zhong et al. [4], Sussna [5], and Wu and Palmer [6]. The main problem of the proposed approaches is that each similarity measure is tied to a particular application or assumes a particular domain model.

1) The measure of Rada et al.

Rada et al. [3] defined a similarity measure for semantic networks based on taxonomic links "*is-a*". To compute the similarity between two ontology concepts, we calculate the distance between them, denoted as dist *C*1, *C*2, in terms of the minimum number of edges which separate them. The similarity measure is defined by the following formula:

doi:10.4304/jetwi.5.4.333-342

$$Sim_{Rada} = \frac{1}{1+\text{dist}_{\text{RADA}}(C1,C2)} \quad (1)$$

with $dist_{RADA}(C1,C2) = len(C1,C2)$ and $len(C1,C2)$ is the length of the shortest path between $C1$ and $C2$. Despite its simplicity, the distance of Rada does not take into account the positions of edges in the concept hierarchy. However, this information influences on the semantic weight of an edge [7].

2) The measure of Zhong

Zhong et al. [4] defines the similarity between two concepts $C1$ and $C2$ by computing the distance between them. This distance is calculated by the positions of the concepts $C1$ and $C2$ in the hierarchy. The model proposed by [4] implies two assumptions: the semantic differences between upper level concepts are bigger than those between lower level concepts (i.e. two general concepts are less similar than two specialized ones) and that the distance between brothers is greater than the distance between parent and child. The similarity measure of [4] is defined as:

$$Sim_{zhong} = 1 - dist_{zhong}(C1,C2) \quad (2)$$

Zhong defines a score (milestone) for every node in the hierarchy obtained from the following formula:

$$milestone(n) = \frac{1/2}{kl^{(n)}} \quad (3)$$

where k is a predefined parameter that enables to intensify or to decrease the speed of evolution of the score according to the depth (k is set to 2 as used in Corese http://www-sop.inria.fr/edelweiss/software/corese/) and $l(n)$ is the depth of the node n in the hierarchy. The distance between two concepts $C1$ and $C2$ is then defined by the milestones of the latters and their closest common parent $ccp(C1,C2)$ as follows:

$$dist_{zhong}(C1,C2) = dist_{zhong}(C1,ccp) + dist_{zhong}(C2,ccp) \quad (4)$$

with

$$dist_{zhong}(C,ccp) = milestone(ccp) - milestone(C)$$

3) The measure of Sussna

The approach of [5] is based on the following idea: "*Let two pairs of concepts separated by the same number of edges (i.e. same length of the shortest path). Then concepts of the deepest pairs (i.e. the furthest away from the root) are closest semantically*". One thus concludes that even with fix distance in the graph, the semantic distance can change. This assumption is justified by the fact that the deeper a node is, the more it is specialized, thus the more it is representative of a precise notion. The distance formula of Sussna is then based on the depth of nodes in the hierarchy and the distance in terms of nodes number.

Besides, [5] seeks to differentiate the different types of relation. For each relation r, the author attributes a weight or a range $[min_r; max_r]$ of weights according to the type of relations that it represents. For example, relations such as hypernymy, hyponymy, holonymy, and meronymy have weights between $\min_r = 1$ and $\max_r = 2$; for antonymy relation, $\min_r = \max_r = 2.5$; and for synonymy, $\min_r = \max_r = 0$. The weight of each edge of type r from some node $C1$ is reduced by a factor which depends on the number of edges, $edges_r^t$, of the same type leaving $C1$:

$$w(C1 \rightarrow r) = \max_r = \frac{\max_r - \min_r}{edges_r(C1)} \quad (5)$$

where $edges_r(C1)$ is the function that computes the number of edges of type r leaving $C1$. It's important to note that Sussna considers that the relations between concepts are not symmetric. In the majority of cases, two opposite relations do not have the same weight (or the same range of weights) and thus if r' is the opposite of r, $w(C1 \rightarrow_r C2) \neq w(C2 \rightarrow_{r'} C1)$. From weights, Sussna defines the distance between two adjacent concepts $C1$ and $C2$ as:

$$dist_S(C1,C2) = \frac{w(C1 \rightarrow_r C2) + w(C2 \rightarrow_{r'} C1)}{2 \times \min[depth(C1), depth(C2)]} \quad (6)$$

The semantic distance between two arbitrary concepts $C1$ and $C2$ is the sum of the distance between the pairs of adjacent nodes along the shortest path connecting them:

$$dist_{Sussna} = \sum_{(x',y') \in sp(C1,C2)} dist'_{Sussna}(x',y') \quad (7)$$

Although this formula employs in theory different types of relations, it was not validated on this point in practice. Moreover, from the version 1.5 of WordNet [8] this formula is not effective any more. WordNet (Fellbaum, 1998) is a large lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (or synsets). Each synset represents a distinct concept. The synsets are connected by relations: synonymy (i.e. words that denote the same concept and are interchangeable in many contexts, all the components of a synset are synonyms), hyperonymy (i.e. *is-a* relation), hyponymy (i.e. the reverse relation of hyperonymy or subsumption relation), meronymy (i.e. *part-whole* or *part-of* relation), holonymy (i.e. the reverse relation of meronymy or *has-a* relation), antonymy (i.e. the *complement-of* relation for the opposites). Nevertheless, the measure of [5] remains interesting since it is the first which introduced the idea that depth plays a main role in the distance.

4) The measure of Wu and Palmer

The formula proposed by [6] computes the similarity between two concepts in an ontology restricted to taxonomic links and it is close to the idea of Zhong regarding the use of the closet common parent of both concepts and their depth in the hierarchy. The similarity between $C1$ and $C2$ is defined by the following formula:

$$Sim_{W\&P}(C1,C2) = \frac{2depth(C)}{depth(C1)+depth(C2)} \quad (8)$$

where C is the most specific common subsumer of $C1$ and $C2$, $depth(C)$ is the number of arcs that separates C from the root of the taxonomy, and $depth(Ci)$ is the number of edges that separates the concept Ci from the root via C.

As a conclusion of this section, the semantic similarity measures presented above have the advantages to be easy to implement. However, they do not take into account the information content of concepts. In what follows, we present some information content approaches.

B. Information Theory or Node-based Approaches

In information theory-based approaches, similarity measures employ a corpus with an ontology restricted to hierarchical links and rest on the computation of the information content IC which a concept represents. This weight must be recalculated with each change of the knowledge base. The notion of information content was first introduced by [1] and was measured by the negative log likelihood of the probability of the concept:

$$IC_{\mathrm{Re}snik(C)} = -\log(P(C)) \quad (9)$$

where $P(C)$ denotes the occurrence probability of concept C in a corpus as well as concepts which it subsumes i.e. (its descendants). Concept frequencies used to estimate concept probabilities are obtained by statistically analyzing a corpus. The probability of encountering an instance of concept C is calculated by the following formula:

$$P(C) = \frac{\sum_{w \in words(C)} freq(w)}{N} \quad (10)$$

where $words(C)$ is the set of words (nouns) subsuming the concept C and N is the total number of words present in the corpus. The idea behind the use of the log function is the more probable a concept is, the less information it expresses. In other terms, frequent words are less informative of infrequent ones. The major disadvantage of the measure of information content lies in the obligation to have a corpus to calculate probabilities. Others authors proposed further measures to compute the information content value of a concept, such as [9] based on the depth and the density and [4] based on the depth.

1) The measure of Resnik

Resnik [1] proposed an alternative way to evaluate the semantic similarity in a taxonomy based on the notion of information content which considers the most informative class instead of the path length. In particular, [1] defines the similarity between two concepts by calculating the information that they share in common. The hypothesis of Resnik is that if two concepts are semantically close, then their closet common parent is close to them and thus its information content is a good indicator. The information shared by two concepts is indicated by the information content of their most specific common subsumer (*mscs*). Accordingly, the information content of a concept C is thus the negative log likelihood. As the probability of encountering the concept C increases, its information content value decreases. The similarity measure of Resnik is then defined as:

$$Sim_{\mathrm{Re}snik}(C1,C2) = IC(mscs(C1,C2)) \quad (11)$$

where $IC(mscs(C1,C2))=-log(P(mscs(C1,C2))$ and $P(mscs(C1,C2))$ denotes the probability of the most specific concept subsumer. We can observe that the higher the position of the *mscs* of both concepts in the hierarchy, the lower their similarity is. If the taxonomy has a unique top node, its probability will be 1, so if the *mscs* of two concepts is the top node for example, their similarity is $-log(1) = 0$. The main limit of Resnik's measure is that it does not take into account the information content of concepts $C1$ and $C2$. Besides, it does not consider the length of the path from the root node to this *mscs* and the depth of concepts $C1$ and $C2$ [7].

2) The semantic similarity of Seco et al.

Seco et al. [9] proposed another measure of the Information Content value which completely rests on the taxonomic structure of WordNet [8]. The assumption behind their method is that concepts with many hyponyms are less informative than concepts that are leaf nodes. In this method, the IC value of a concept depends on the number of its hyponyms and a constant.

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

where $hypo(C)$ is the function which returns the number of hyponyms of a given concept and max_{wn} is a constant that is set to the maximum number of concepts existing in the taxonomy. To evaluate their IC metric, [9] compared the results of the similarity measures of Resnik [1], Lin [10], and Jiang and Conrath, [11] when using the IC value of [1] with those when using their IC value by correlating the similarity scores with those of human judgments provided by Miller and Charles [12]. The evaluation confirmed the authors' initial assumption regarding the usefulness of the hierarchical structure and suggests the use of other taxonomies such as Gene Ontology (http://www.geneontology.org/) to have a generalized metric and thus achieving domain independence.

3) The universal similarity measure of Lin

Lin [10] tried to define a universal similarity measure that would be applicable to different domains (e.g. ordinal values, feature vectors, word similarity, and semantic similarity in a taxonomy) or knowledge representation forms. This measure was derived from a set of assumptions and captures the following three intuitions about similarity:

1. The similarity between two objects *A* and *B* is related to their commonality; the more commonality they share, the more similar they are.
2. The similarity between two objects *A* and *B* is related to the difference between them; the more differences they have; the less similar they are.
3. The maximum similarity between two objects *A* and *B* is reached when *A* and *B* are identical, no matter how much commonality they share.

Lin [10] defined the commonality between *A* and *B* as the information content of the proposition that states the commonalities between them:

$$I(common(A,B)) \tag{13}$$

Besides, Lin defined the difference between *A* and *B* as:

$$I(description(A,B)) - I(common(A,B)) \tag{14}$$

where *description(A,B)* is a proposition describing what *A* and *B* are. Based on these assumptions, [10] proved the following similarity theorem: The similarity between *A* and *B* is measured by the ratio between the amount of information needed to state their commonality and the information needed to fully describe what they are:

$$Sim_{Lin}(A,B) = \frac{\log P(common(A,B))}{\log P(description(A,B))} \tag{15}$$

According to [10], it is only necessary to specify the probability computation according to a domain in order to obtain its own similarity measure. To demonstrate this assumption, [10] provides as examples, a similarity between strings, between two words based on a corpus, between two concepts of a taxonomy, and between ordinal values [7]. The similarity measure that Lin [10] proposed between two concepts $C1$ and $C2$ in a taxonomy is expressed by:

$$Sim_{Lin}(C1,C2) = \frac{2\times \log P(mscs(C1,C2))}{(\log P(C1) + \log P(C2))} \tag{16}$$

where the probabilities $P(C)$ are obtained w.r.t Resnik's $P(c)$ (10). In this measure, [10] defined the shared information content between two concepts $C1$ and $C2$ by 2× the information content of their most specific common subsumer ($mscs(c1, c2)$) and the information content of the description by the sum of the descriptions of the two objects. To evaluate his similarity measure, Lin computed the similarity between 28 pairs of concepts taken from WordNet using his measure and those of Resnik [1] and Wu and Palmer [6] and correlated the obtained scores with scores assigned by human subjects in the experiments of Miller and Charles. The comparison shows that his similarity measure presents a slightly higher correlation with human judgments than the other two measures [10].

C. Hybrid Approaches

Hybrid approaches combine edge-based techniques and information content by considering the shortest path between two concepts and the density of all nodes along this same path in the similarity computation. Information content values are obtained through statistical analysis of corpora and are taken into account as a decision factor.

1) The measure of Jiang and Conrath

The measure of Jiang and Conrath [11] combines the information content of the most specific common subsumer and those of the concerned concepts, and consequently, it can mitigate the limits of Resnik's method. Their semantic similarity relies on the importance degree of a link in the graph and the local density of a node, its depth and type [7]. Recalling the definition of information content, they defined the strength of a link as:

$$dist_{J\&C}(C,parent(C)) = IC(C) - IC(parent(C)) \tag{17}$$

Besides, recalling the idea of the shortest path between two concepts in the taxonomy, the semantic distance of Jiang and Conrath between an arbitrary pair of concepts is given by the sum of distances along the shortest path that connects these concepts:

$$\mathrm{dist}_{J\&C}(C1,C2) = \sum_{c\in sp(C1,C2)} \mathrm{mscs}(C1,C2)\,\mathrm{dist}_{J\&C}(C,\mathrm{parent}(C)) \tag{18}$$

where $sp(C1,C2)$ denotes the set of all nodes in the shortest path from $C1$ to $C2$. The node $mscs(C1,C2)$ is removed from $sp(C1,C2)$ in this formula because it has no parent in the set. Considering (17) and (18), the final Jiang and Conrath's semantic distance formula between two concepts $C1$ and $C2$ is defined as:

$$dist(C1,C2) = (IC(C1)+IC(C2)) - (2?IC(mscs(C1,C2))) \tag{19}$$

This distance contains the same components as the Lin's similarity however their combination is not a ratio but a difference. The similarity measure of [11] is then defined by the reverse of the semantic distance:

$$Sim_{J\&C}(C1,C2) = \frac{1}{dist(C1,C2)} \tag{20}$$

2) The similarity measure of Leacock and Chodorow

The similarity measure of Leacock and Chodorow [13] takes into account the path length between concepts in an ontology restricted to taxonomic links and the depth of the taxonomy:

$$Sim_{L\&C}(C1,C2) = -\log\left(\frac{len(C1,C2)}{2譒AX}\right) \tag{21}$$

where $len(C1,C2)$ is the length of the shortest path between two concepts $C1$ and $C2$ and MAX is the maximum taxonomy depth of the information source. The path length is measured by the number of nodes in the path instead of links. The hypothesis of Leacock and Chodorow is to approximate the probability by taking

into account the path length. The measure of [13] enables to avoid the computation of the information content but it keeps the concept of the information theory. It transforms the Rada distance into a similarity. In the same way, measures which consider only the shortest path length are imprecise because they do not take into account the density or depth of concepts.

IV. Semantic Relatedness Measures

Semantic relatedness measures [7, 14-16] compute the degree to which a pair of concepts are related considering the whole set of semantic links among them. Consequently, semantic relatedness is a generalization of semantic similarity. In other terms, similar concepts are also semantically related but the inverse is not necessarily true, that is, concepts which are related by lexical or functional relationship can be dissimilar. The computation of semantic relatedness has many applications in different areas, such as natural language processing, information extraction and retrieval, lexical selection, automatic correction of word errors in text, and word sense disambiguation. In this section, we present some methods which have been proposed to compute degrees of relatedness among texts, words or concepts. These measures can be classified into lexical resource-based measures, Wikipedia-based measures, and Web-based measures according to the source of knowledge utilized.

A. The Relatedness Measure of Hirst and St-Onge

In [14], Hirst and St-Onge proposed a WordNet-based definition of semantic relatedness which seeks relation between two different words considering their synsets. In particular, they defined three types of relation between two words: extra-strong, strong and medium-strong. A relation between two words is strong if: (a) they have a synset in common (e.g. human and person), (b) they are associated to different synsets interlinked by an horizontal link (e.g. precursor and successor), or (c) there is any type of link between a synset associated with each word and one of the words is a compound word that includes the other (e.g. school and private school). A relation between two words is medium-strong if there is an allowable path connecting synsets of the related words. A path is allowable if it does not contain more than five links between synsets and respects one of the eight allowable patterns. The hypothesis behind this is "*The longer the path and the more changes of direction, the lower the weight*"[14]. The authors have associated a direction among the values *Upward* (i.e. a generalization link), *Downward* (i.e. a specialization link), and *Horizontal* (i.e. antonymy or similarity links) for each relation type in WordNet. The directions are assigned according to the type of links in WordNet. The allowable eight patterns of paths in a medium-strong relation are U, UD, UH, UHD, D, DH, HD, H. In this method, the similarity computation is based on the allowable patterns of path. Once a regular path is found, the weight of the relation type (i.e. the path between two words) is defined by:

$$\mathrm{Rel}(C1,C2)=\begin{cases}3\times \mathrm{C(extra-strong)}; 2\times \mathrm{C(strong)}\\ \mathrm{C-len}(C1,C2)-k\times \mathrm{turns}(C1,C2)(\mathrm{medium-strong})\end{cases} \quad (22)$$

The semantic relatedness of [14] is defined by:

$$Sim_{H\&S} = weight(path(C1,C2)) = C - len(C1,C2) - k\ ?tu\ rns(C1,C2) \quad (23)$$

where C and k are two constants (they are fixed to C = 8 and k = 1 [2]), $len(C1,C2)$ is the length of the shortest path taking into account the directions that are affected according to the type of relation in WordNet, and $turns(C1,C2)$ is the number of direction changes in the path. This measure adapts the Rada's measure to further take account of non-hierarchical relations in an ontology. Thus, it has the same limits of the Rada's measure as it does not consider the density or the depth of concepts and it does not make use of information contents that nodes represent (i.e. it assumes that the information content of all nodes is uniform).

B. The Relatedness Measure of Mazuel and Sabouret

Mazuel and Sabouret [7] focused on the issue of semantic relatedness in a semantic network and proposed a new semantic distance to compute the degree of relatedness between two concepts of a taxonomy augmented with non-hierarchical relations. This measure takes into account different kinds of relations (i.e. subsumption (is-a), meronymy (part-of) or any other domain specific relation) and uses a set of rules to discard unallowable paths generated by the presence of non-hierarchical relations. These rules are inspired from the works of the patterns of semantically correct paths of [14]. In this method, [7] distinguished between single-relation paths and multiple-relation paths and proposed measures for each situation. A single-relation path is a path whose edges are all of the same type: a hierarchical path (i.e. representing the relation *is-a*) or a non-hierarchical path. To compute the weight of a hierarchical single-relation path between two concepts x and y in the ontology, [7] reused the Jiang and Conrath measure which is given by the difference between weights of the two concepts:

$$W(path_{X\in\{is\text{-}a,includes\}}(x,y)) = |IC(x) - IC(y)| \quad (24)$$

To compute the weight of a single-relation path when the relation is not hierarchical, [7] proposed a new formula because the information content of nodes is calculated according to the hierarchical structure of the taxonomy. The authors associated a static weight to each relation type that represents its semantic cost. Besides, they based their formula on the n/n+1 function which simulates the log form. Consequently, the weight of a path between two concepts x and y given its weight and its length is defined by:

$$W(path_X(x,y)) = TC_X \times \left(\frac{|path_X(C1,C2)|}{|path_X(C1,C2)|+1}\right) \quad (25)$$

In the case of a mixed-relation path which contains different kinds of relations, [7] proposed to decompose it in an ordered set of *n* sub-paths based on the transitive nature of the edges of a single-relation path. The weight of a mixed-relation path between two concepts *x* and *y* is defined as the sum of weights of sub-paths composing the minimal decomposition of the path:

$$W(path(x,y)) = \sum_{p \in T_{\min}(path(x,y))} W(p) \quad (26)$$

where $T_{min}(path(x, y))$ is the unique ordered set of sub-paths. Besides, [7] demonstrated that the weight of an hierarchical mixed-relation path containing only two kinds of relations: (a) the relation *is-a* between concept *C1* and the mscs of concepts *C1* and *C2*, and (b) the relation *includes* from the mscs to concept *C2* corresponds to the Jiang and Contrath's distance:

$$\begin{aligned} &W(path(C1,C2)) = W(path_{is\text{-}a}(C1,mscs(C1,C2))) \\ &+W(path_{includes}(mscs(C1,C2),C2))) \\ &=|IC(C1)- IC(mscs(C1,C2))|+|IC(mscs(C1,C2))- IC(C2)| \\ &=IC(C1)+ IC(C2)- 2\ ?IC(mscs\ (C1,C2)) \end{aligned} \quad (27)$$

The final distance measure of [7] considers only the semantically correct paths between two concepts *C*1 and *C*2, and thus it corresponds to the minimal weight among the set of valid paths as defined in the following formula:

$$dist(C1,C2) = \min_{\{p \in \Pi_{(C1,C2)} | HSO(p)=true\}} W(p) \quad (28)$$

where $\Pi(C1,C2)$ is the set of elementary paths (i.e. acyclic) between the concepts *C1* and *C2*. To obtain the set of valid paths, the authors used the function $HSO : \Pi(C1,C2) \rightarrow B$ which determines, according to the path patterns of Hirst and St-Onge [14], if a path is semantically correct (i.e. *HSO*(*p*) = *true*) or not (i.e. *HSO*(*p*) = *false*). This distance can be converted in a semantic relatedness measure by using the classic linear conversion of Resnik:

$$rel(C1,C2) = 2\ ?IC_{\max} - dist(C1,C2) \quad (29)$$

To evaluate their semantic relatedness measure, [7] employed two testing sets from the literature: the test of Miller & Charles [12] and the test of WordSimilarity-353 [17]. Besides, they compared their measure with the similarity measures of Rada, Resnik, Lin and Jiang Contrath and the relatedness measure of Hirst St-Onge. Experimental results show that they obtained best correlation w.r.t human judgments. However, [7] considered only the noun sub-part of WordNet 3.0 and the test focused only on the non-hierarchical transitive relation "*part-of*". In addition, this measure rests on a taxonomic model augmented with one heterogeneous relation and needs to be extended to model complex relations between concepts (i.e. intersections, disjunctions of classes, etc.) such as OWL-Lite .

C. Web-based Semantic Relatedness Measure of Gracia and Mena

In [15], Gracia and Mena proposed the *NormalizedWebDistanceNWD(x,y)* which is a generalization of the Cilibrasi and Vitanyi's Normalized Google Distance NGD(x, y) (30) to compute semantic relatedness between two plain words (or search terms) indexed by different Web search engines.

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (30)$$

where $f(x)$ denotes the number of pages containing *x*, $f(x, y)$ denotes the number of pages containing both words *x* and *y*, and *N* is a normalizing factor. Frequencies are computed using Google page counts. The proposed semantic relatedness measure between two search terms *x* and *y* is defined as:

$$relWeb(x,y) = e^{-2NWD(x,y)} \quad (31)$$

In a later version of their work, Gracia and Mena have taken the word relatedness as a basis to define a new measure that computes how much a pair of ontology terms are semantically related. This measure captures the following desirable features:

- **Domain independent:** it computes relatedness between terms from different ontologies by exploiting some elements of their available semantic descriptions.
- **Universality:** it does not rely on specific lexical resources (e.g. corpus, dictionaries, or WordNet) or knowledge representation languages (e.g. OWL).
- **Maximum coverage:** since it uses the Web as knowledge source, it guarantees a maximum coverage of possible interpretations of the words and thus it extends the scope of applications (e.g. word sense disambiguation, ontology matching, etc.).

To do that, the proposed method computes the degree of semantic relatedness between a pair of senses that two ontological terms represent (i.e. a class, a property or an instance) by considering two levels of semantic description: Level0 which represents the term label and its synonyms and Level1 which represents the ontological context of the term. This latter describes the set of other ontological terms and it corresponds to (a) the set of direct hypernyms if the term is a class, or (b) the set of domain classes if the term is a property, or (c) the class it belongs to if the term is an instance. The relatedness between two ontological terms *a* and *b* at Level0 is computed by (32) whereas the relatedness at Level1 is measured by (33).

$$rel_0(a,b) = \frac{\sum_{i,j} relWeb(syn_{ai}, syn_{bj})}{|Syn(a)|.|Syn(b)|} \quad (32)$$

$$i = 1..|Syn(a)|$$

$$j = 1..|Syn(b)|$$

where *Syn*(*a*) and *Syn*(*b*) denotes the set of synonyms of terms *a* and *b* and *OC*(*a*) and *OC*(*b*) denotes their ontological context.

$$rel_1(a,b) = \frac{\sum_{i,j} rel_0(OC_{ai}, OC_{bj})}{|OC(a)|.|OC(b)|} \quad (33)$$

$$i = 1..|OC(a)|$$
$$j = 1.|OC(b)|$$

The final relatedness degree between two ontological terms is the combination of the semantic relatedness values obtained from (32) and (33) after being weighted as follows:

$$rel(a,b) = w_0.rel_0(a,b) + w_1.rel_1(a,b) \quad (34)$$

where $w_0 \geq 0, w_1 \geq 0$ and $w_0 + w_1 = 1$. Gracia and Mena [15] considered only two levels of semantic description for a term based on an assumption derived from Resnik's idea [1] which assumes that the higher a word is in the hierarchy that characterize the sense of an ontological term, the lesser information content it expresses, and consequently it is less significant to characterize the term. As future works, [15] planned to explore other variations of the method by weighting differently the synonyms of a term in (32) or by considering alternative definitions of the ontological context. Finally, the authors proposed a mixed relatedness measure between ontology terms and plain words in order to cover other usage scenarios. Relatedness at *Levels* 0 and 1 are computed by the following equations:

$$rel_0(t,w) = \frac{\sum_i relWeb(syn_{ti}, w)}{|Syn(t)|} \quad i = 1..|Syn(t)| \quad (35)$$

$$rel_1(t,w) = \frac{\sum_i rel_0(OC_{ti}, w)}{|OC(t)|} \quad i = 1..|OC(t)| \quad (36)$$

The final relatedness between an ontology term *t* and a plain word *w* is a combination of results of (35) and (36):

$$rel(t,w) = w_0.rel_0(t,w) + w_1.rel_1(t,w) \quad (37)$$

where $w_0 \geq 0$, $w_1 \geq 0$, and $w_0 + w_1 = 1$. Besides, the authors carried out two experiments to test the application of their Web-based relatedness measure in disambiguation and ontology matching tasks. Experiments were done using a new test data set consisting of 30 pairs of English nouns that are connected with different types of relations (e.g. similarity, meronymy, frequent association, etc.) and rated for semantic relatedness by a group of 30 university graduated persons on a scale ranging from 0 to 4 (i.e. from no relatedness to identical or strongly related words). They used the Spearman's rank correlation coefficient to determine the correlation between their results and those of humans. The results show that Web-based measures present a better correlation with human judgments than WordNet-based measures.

D. The Extended Gloss Overlap Measure of Banerjee and Pedersen

Banerjee and Pedersen [18] proposed another measure to quantify semantic relatedness between concepts, namely, the extended gloss overlap measure, which is based on the computation of the number of shared words (or overlaps) in the concepts definitions taken from a machine readable dictionary. The basic idea of this approach consists in expanding the glosses of the words being compared by including also glosses of concepts which are recognized to be related to them and their neighbors according to explicit relations provided in the lexical database WordNet. This approach extends the one proposed by Lesk [19] who assumed that related word senses are usually described using the same words and thus he defined a relatedness measure based on gloss overlaps but which considers only overlaps among the glosses of the candidate senses of the target word and those that surround it in the given context. This is a considerable limitation as most dictionary glosses tend to be short and therefore do not provide enough words to find overlaps with. The proposed measure (38) takes as input a pair of synsets and generates a numeric value of semantic relatedness based on the number of overlapping words in their respective glosses as well as in the glosses of synsets they are connected to in a given concept hierarchy. In order to test the proposed relatedness measure, [18] developed an approach to word sense disambiguation (WSD) task which assigns a sense to a target word in a given context that is the most related to the senses of its neighbors using this measure. Evaluation of the measure based on the approach of comparison to human judgments showed a satisfactory correlation coefficient, but word sense disambiguation experiments showed that considering extended gloss overlaps improves the disambiguation results and yields much better than the original Lesk Algorithm [19]. The authors plan to augment the scores of overlaps with global statistics about the word occurrences and to evaluate the measure on different NLP tasks. The relatedness score between the inputs synsets *A* and *B* is measured as the sum of scores of phrasal gloss overlaps between them:

$$relatedness(A,B) = \sum_{\forall (R1,R2) \in RELPAIRS} score(R1(A), R2(B)) \quad (38)$$

where *RELPAIRS* denotes the set of all possible relation pairs formed from the set of relations defined in WordNet (e.g. hypernyms, hyponyms, meronyms, holonym, also-see relation, attribute, pertainym), and *score*() is the function which detects and scores the phrasal gloss overlaps between the inputs. The scoring mechanism consists in assigning a phrasal *n* word overlap the score of n^2.

V. Semantic Measures Evaluation

From the literature, a way to evaluate the results of semantic similarity measures is to find a good correlation between the computed similarity scores and the average similarity ratings provided by human evaluators in benchmarks, such as, Miller and Charles [12] and WordSimilarity-353 [17]. The higher the correlation of a method, the better the method is, i.e. the more it approaches the results of human judgments. A correlation is a number between -1 and +1 which measures the degree of relationship between two variables. A positive value for the correlation implies a positive association whereas a negative value implies a negative or inverse association. The two most commonly used measures of correlation are the Pearson's correlation coefficient and the Spearman's rank correlation coefficient. The Pearson's correlation coefficient enables to analyze linear relations between two variables based on their actual values. The correlation coefficient ranges between -1 and +1 and it is interpreted as follows:

- Near to -1: the two vectors are opposite or negative agreement/disagreement.
- Around 0: the two vectors are independent or no agreement.
- Near to 1: the two vectors are dependent positive agreement.

Let two vectors of length $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, the Pearson's correlation coefficient is defined as follows:

$$p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (39)$$

The Spearman rank correlation coefficient (40) is a non-parametric measure of correlation which uses ranks to calculate the correlation rather than absolute values. The correlation coefficient is a number ranging also between -1 (total disagreement) and +1 (total agreement). A positive correlation is one in which the ranks of both variables increase together. A negative correlation is one in which the ranks of one variable increase as the ranks of the other variable decrease. A correlation of +1 or -1 will arise if the relationship between the two variables is exactly linear. A correlation close to zero means that there is no linear relationship between the ranks.

$$p = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \quad (40)$$

Some software packages were already proposed that enable to compute similarity, such as the WordNet::Similarity package (http://talisker.d.umn.edu/cgi-bin/similarity/similarity.cgi) and the Nuno Seco package [9]. The WordNet::Similarity package consists of Perl modules that implement the following WordNet-based measures: Resnik [1], Lin [10], Jiang and Conrath [11], Leacock and Chodorow [13], Hirst and St-Onge [14], Wu and Palmer [6], the extended gloss overlaps measure of Banerjee et al. [18], and two measures based on context vectors by Patwardhan and Pedersen [20]. The Nuno Seco package is implemented in java and can be downloaded from the rubric "extension" in WordNet site. In what follows, we present the most commonly used similarity benchmarks, namely, Miller and Charles [12] and WordSimilarity-353 [17].

A. Benchmarks for Semantic Measures Evaluation

1) Miller and Charles test

The test of Charles and Miller [12] is a set of 30 pairs of nouns with their similarity ratings determined by human judgments. 38 undergraduate students have participated to the test and were asked to rate the similarity of each pair on a scale from 0 (not similar) to 4 (perfect synonymy). The average rating of each pair represents a good estimate of how similar the two words are. Most of works presented in the previous section end with an evaluation w.r.t 30 pairs selected by Miller and Charles. This protocol enabled to fix a work base for the research community on semantic distances. In fact, if a distance measure reached the coefficient of 0.91, then it will be representative of the real distance of a human judgment. However, the test of Miller and Charles is based on synonymy judgment, thus, it is mainly oriented to evaluate similarity measures not relatedness measures. Most of selected word couples do not have functional relations between them since it was explicitly requested to human subjects to judge similarity between concepts. Hence, the dataset of Miller and Charles is not adapted to test functional relations between two concepts and accordingly to evaluate a semantic relatedness measure.

2) The WordSimilarity-353 test collection

The WordSimilarity-353 benchmark [17] is a set of 353 English word pairs for which subjects were asked to estimate the similarity or relatedness of words on a scale from 0 (totally unrelated words) to 10 (strong related or identical words). The WordSimilarity-353 test set can be used to train and to test algorithms implementing semantic measures. It was proposed in order to mitigate the problems of Miller and Charles test. It includes all the 30 noun pairs of Miller and Charles. Agirre et al. [21] proposed to split the WordSimilarity-353 dataset into two subsets (http://alfonseca.org/eng/research/wordsim353.html), the first subset contains the union of similar and unrelated pairs and focuses on computing similarity whereas the second subset contains the union of related and unrelated pairs and focuses on computing relatedness.

As a conclusion, human judgments of similarity and relatedness provided in benchmarks presented above are supposed to be correct by definition and give clear evaluation of the performance of a measure. However, the main drawback of this approach lies in the difficulty of obtaining a large set of reliable and subject-independent judgments for comparison.

B. Approaches for Semantic Measures Evaluation

In [2], Budanitsky and Hirst focused on comparing the performance of five WordNet-based measures (Hirst and

St-Onge, Jiang and Conrath, Leacock and Chodorow, Lin, and Resnik) based on two evaluation approaches: comparison of computed semantic relatedness or similarity scores with human judgments and comparison of the performance of these measures in a particular application. To compute the frequency of concepts needed in the information-theoretic approaches, [2] used the Brown Corpus of American English [22]. For the first evaluation method, [2] computed the correlation coefficients between human and computer ratings of the word pairs of Rubenstein-Goodenough and Miller-Charles in order to determine the strength of the linear association between them. The comparison has shown that the difference between the values of the highest and lowest correlation coefficients for the test of Miller and Charles and the test of Rubenstein-Goodenough are in the order of 0.1 and 0.06 respectively. Besides, [2] employed the upper bound for the Miller and Charles word pairs to compare the performance of the selected measures on it and they found that the correlation coefficients compare quite favorably with this upper bound. Moreover, [2] concluded that the measures do not react in the same way toward the increase of the size of the dataset. In fact, while the correlation coefficients with human judgments of *relH&S*, *simL&C* and *simR* improve, those of *distJ&C* and *simLin* deteriorate.

As [2] point out, though human evaluation approach is considered as the best method to evaluate a similarity or a relatedness measure, its main drawback stands for the difficulty of acquiring considerable amounts of test sets of word pairs with human-assigned scores. Besides, [2] continue to add that they need in NLP tasks human judgments of the relatedness of word-senses not just words. This need can be satisfied by exploring contexts. In order to overcome the problems posed by this approach, [2] used an application-based evaluation approach which compares relatedness measures based on their ability to detect and correct semantic anomalies such as malapropisms. To test the measures, they created a corpus of malapropisms. Then, they tried to detect and correct them by an algorithm that uses the five measures of semantic relatedness with different searching scopes. They considered it as a retrieval task and evaluated it in terms of Precision, Recall, and F-measure. The analysis of differences between measures' results for the malapropism suspicion phase shows that the Jiang and Conrath's measure outperforms the others in all scopes. The results for malapropism detection phase shows also that the measure of Jiang and Conrath does better than the other measures. Besides, evaluation shows that though the measure of Hirst and St-Onge is the only one among the others that focuses on computing semantic relatedness, it presents poor performance in both stages. To support the evaluation results of [2] regarding the performance of the Jiang and Conrath's measure, we considered other approaches in the NLP domain that also apply WordNet-based measures and we perceived that these results are consistent with the experiments' results of approaches of Stevenson and Greenwood [23], Kohomban and Lee [24], and Patwardhan et al.[25]. In fact, [23] proposed a semantic similarity approach to information extraction pattern acquisition which relies on comparing patterns similarity using their own measure. This measure takes into account pattern vectors and their transposes and a similarity matrix which contains information about semantic similarity between pairs of lexical items. They experimented several measures in order to populate the semantic similarity matrix and found that the measure defined by Jiang and Conrath is the most effective one. Similarly, [24] described a method to learn generic semantic classes of a given word instance in order to mitigate the lack of training data problem in word sense disambiguation. In this method, [24] computed the relatedness between the sense of the test word and the most frequent sense of it within the candidate class using different similarity measures. The experiments showed that the measure of Jiang and Conrath gives best results for this task. In the same way, [25] carried out word sense disambiguation experiments to evaluate the same five measures of semantic relatedness that have been also compared by [2] in addition to the extended gloss overlap measure. Experiments were performed using noun data gathered from the English Lexical sample task of SENSEVAL-2 (http://www.senseval.org/). Similarly, the authors found that the extended gloss overlap measure of Banerjee and Pedersen [18] and the semantic distance measure of Jiang and Conrath [11] result in the highest accuracy.

VI. Conclusion

In this paper, we presented a classification of semantic measures and discussed the basics of the various approaches proposed for each class. The benchmarks and evaluation approaches that are commonly used by researchers to assess the quality of their semantic measure proposals are also stated. This survey could help researchers to choose the most appropriate similarity or relatedness measure for their needs.

References

[1] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th International Joint Conf. on Artificial Intelligence* - Volume 1, 1995, pp. 448–453.

[2] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatednes," *Computational Linguistics*, vol. 32, no. 1, 2006, pp. 13–47.

[3] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE transactions on systems, man, and cybernetics*, vol. 19, no. 1, 1989, pp. 17–30.

[4] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual graph matching for search," in *Proc. 10th International Conf. on Conceptual Structures: Integration and Interfaces*, London, UK: Springer-Verlag, 2002, pp. 92–106.

[5] M. Sussna, "Word sense disambiguation for free-text indexing using a massive semantic network," in *Proc. Second International Conf. on Information and Knowledge Management*, New York, NY, USA: ACM, 1993, pp. 67–74.

[6] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. 32nd Annu. Meeting on Association for*

Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133–138.

[7] L. Mazuel and N. Sabouret, "Semantic relatedness in semantic networks," in *Proc. 18th European Conf. on Artificial Intelligence*, Amsterdam, The Netherlands: IOS Press, 2008, pp. 727–728.

[8] C. Fellbaum, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[9] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in wordnet," in *Proc. 16th European Conf. on Artificial Intelligence*, 2004, pp. 1089–1090.

[10] D. Lin. "An Information-Theoretic Definition of Similarity", in *Proc. of 15th International Conf. on Machine Learning*, SanFrancisco, CA, USA, 199, pp. 296–3048.

[11] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. International Conf. Research on Computational Linguistics*, Taiwan, 1997, pp. 19–33.

[12] G. Miller and W. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, 1991, pp. 1–28.

[13] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet: A Lexical Reference System and its Application*, C. Fellbaum, Ed. Cambridge, Massachusetts: MIT Press, 1998, pp. 265–283.

[14] G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, MIT Press, 1998, pp. 305–332.

[15] J. Gracia and E. Mena, "Web-based measure of semantic relatedness," in *Proc. 9th international conf. on Web Information Systems Engineering*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 136–150.

[16] E. Blanchard, M. Harzallah, H. Briand, and S. Kuntz, "A typology of ontology-based semantic measures," *in Proc. Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability*, 2005.

[17] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: the concept revisited," *ACM Transactions On Information Systems*, vol. 20, no. 1, January 2002, pp.116–131.

[18] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *Proc. 18th International Joint Conf. on Artificial intelligence*, 2003, pp. 805–810.

[19] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in *Proc. 5th Annu. International Conf. on Systems Documentation*, pp. 24–26, 1986.

[20] S. Patwardhan and T. Pedersen, "Using WordNet-based context vectors to estimate the semantic relatedness of concepts," in *Proc. EACL 2006 workshop making sense of sense - Bringing computational linguistics and psycholinguistics together*, 2006, pp. 1–8.

[21] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proc. Human Language Technologies: The 2009 Annu. Conf. of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 19–27.

[22] N. W. Francis and H. Kucera, "Frequency Analysis of English Usage: Lexicon and Grammar," *J. of English Linguistics*, vol. 18, no. 1, 1982, pp. 64–70.

[23] M. Stevenson and M. A. Greenwood, "A Semantic Approach to IE Pattern Induction," in *Proc. 43rd Annu. Meeting on Association for Computational Linguistics*, 2005, pp. 379–386.

[24] U. S. Kohomban and W. S. Lee, "Learning semantic classes for word sense disambiguation," in *Proc. 43rd Annu. Meeting on Association for Computational Linguistics*, 2005, pp. 34–41.

[25] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation," in *Proc. 4th International Conf. on Computational Linguistics and Intelligent Text Processing*, 2003, pp. 241–257.