A typology of ontology-based semantic measures

Emmanuel Blanchard, Mounira Harzallah, Henri Briand, and Pascale Kuntz

Laboratoire d'Informatique de Nantes Atlantique
Site École polytechnique de l'université de Nantes
rue Christian Pauc
BP 50609 - 44306 Nantes Cedex 3 - France
emmanuel.blanchard@univ-nantes.fr

Abstract. Ontologies are in the heart of the knowledge management process. Different semantic measures have been proposed in the literature to evaluate the strength of the semantic link between two concepts or two groups of concepts from either two different ontologies (ontology alignment) or the same ontology. This article presents an off-context study of eight semantic measures based on an ontology restricted to subsomption links. We first present some common principles, and then propose a comparative study based on a set of semantic and theoretical criteria.

1 Introduction

A consensus is now established about the definition and the role of an ontology in knowledge engineering: "an ontology is a formal, explicit, specification of a shared conceptualization" [1], it constitutes knowledge repository that supports information exchanges between computer systems. Several application fields search to exploit their wealth: semantic web, system interoperability, competence management, e-learning, natural language processing, etc. Numerous semantic measures on ontology have been proposed in literature to evaluate the strength of the semantic link between two concepts or two groups of concepts from two different ontologies (ontology alignment) or inside an ontology. The majority of the semantic measures defined on a unique ontology is developed and validated in a specific context. This article presents an off-context study of eight semantic measures whose definitions take only into account subsomption links. The synthesis of the parameters which appear in at least one of these measures has provided us a basis of comparison to develop a semantic measure typology.

First, we identify ontology-based semantic measure characteristics and the parameters that influence these measures. Then, we present a comparative study of eight measures structured according to the previously defined criterion.

2 Semantic measure classification

We consider here the basic primitives of an ontology which are concepts and relations. Among the set of possible relations, some of them are not used systematically. For instance, the taxonomic relations (hyperonymy/hyponymy) which

correspond to the subsomption link (is-a) are the commonly used relations. Additional relations may also appear e.g. partonomic relations (meronymy/holonymy) which correspond to the composition link (part-of), lexical relations (synonymy, antonymy, etc.) [2][3]. We have here decided to focus ourselves on taxonomic relations before we generalize this study to other relation types. This choice is guided by most of previous works which are limited to one or some relation types but which always consider taxonomic relations.

Measure validation is evoked according to three ways [4]: *mathematical analysis*, *comparison with human judgment* and *application specific evaluation*. In this paper we favour the mathematical analysis, and we introduce the different characteristics of semantic measures, and the different parameters which influence them. When defining a measure, three characteristics are generally specified:

Information sources. Each considered measure is based on a given ontology (most often WordNet). Some definitions require a corpus of texts to add information such as the distribution of concept frequencies.

Principles. Most of measures are based on axiomatic principles e.g. they make functions of the information content or the shortest path length.

Semantic class. Different classes have been introduced in the literature: *semantic distance*, *semantic similarity* and *semantic relatedness* between two concepts in the same ontology.

The *semantic similarity* evaluates the resemblance between two concepts from a subset of significant semantic links (e.g is-a and part-of). The *semantic relatedness* evaluates the closeness between two concepts from the whole set of their semantic links. All pairs of concepts with a high semantic similarity value have a high semantic relatedness value whereas the inverse is not necessarily true. The *semantic distance* evaluates the disaffection between two concepts; it is an inverse notion to the semantic relatedness.

We have identified four parameters associated with the ontology taxonomic hierarchy which influence at least one of the measures:

- p_1 *the length of the shortest path*: the length of the shortest path between two concepts c_i and c_j;
- p_2 *the depth of the most specific common subsumer*: the length of the shortest path between the root and the most specific common subsumer of c_i and c_j;
- p_3 *the density of the concepts of the shortest path*: the number of sons of each concept which belongs to the shortest path between two concepts c_i and c_j;
- p_4 *the density of the concepts from the root to the most specific common subsumer*: the number of sons of the concepts which belong to the shortest path from the root to the most specific common subsumer of two concepts c_i and c_j.

Since our study is restricted to the taxonomic hierarchy, p_1 is equal to the sum of the two shortest path from the two concepts to their most specific common subsumer. The parameter p_2 is the difference between the length of the shortest path between the root and one of the two concepts and the length of the shortest path between this concept and the most specific common subsumer.

The measures which also use a corpus consider the information content of some concepts. Let us consider a concept c. The information content is defined by $CI(c) = -\log(P(c))$ where $P(c)$ corresponds to the occurrence probability, in a consequent corpus of texts, of c or one of its directly or indirectly subsumed concepts. Let us notice that this definition contains the information on the shortest path from the root to the concept c (depth of c). As $P(c)$ is an exponential decreasing function of the depth of c, $CI(c)$ is proportional to this latter. In addition, the information extracted from a corpus by this approach contains also the information of the density of the concepts on the same path. Indeed, let us consider the set S of the concepts which have the same father (direct subsumer) as c. When the cardinality of S increases, the average occurrence probability of each element of S decreases. Consequently, $CI(c)$ increases.

3 Semantic measure presentation

In the following, we present eight measures using information sources and principles defined in the previous section. Then, we analyse the theoretic definition of each measure. Some of the following ontology or corpus characteristics are considered in the definitions:

rt: ontology root
$pths(x, y)$: set of paths between the concepts x and y
$len_e(x)$: length in number of edges of the path x
$len_n(x)$: length in number of nodes of the path x
$P(x)$: occurrence probability of a concept x in a corpus
$mscs(x, y)$: the most specific common subsumer of x and y
$trn(x)$: number of direction changes of the path x
min_r: minimum weight assigned to the relation r
max_r: maximum weight assigned to the relation r
$n_r(x)$: number of relations of type of r which leave x

Rada et al.'s distance[5]. It is based on the shortest path between two concepts c_i and c_j (p_1) in an ontology restricted to taxonomic links:

$$dist_{rmbb}(c_i, c_j) = \min_{p \in pths(c_i, c_j)} len_e(p) \tag{1}$$

When considering the shortest path, all the taxonomic links between two adjacent concepts are supposed to have a same value.

Resnik's similarity[6]. It is established on the following hypothesis: the more the information two concepts share in common, the more similar they are. Like the previous measure (1), this one only considers taxonomic links. On the basis of the information theory, Resnik proposes to add the information content. The information shared by two concepts is indicated by the information content of their most specific common subsumer:

$$sim_r(c_i, c_j) = -\log P(mscs(c_i, c_j)) \tag{2}$$

The use of the information content of the most specific common subsumer implies that this measure depends on two parameters: the length of the shortest path from the root to the most specific common subsumer of c_i and c_j (p_2) and the density of concepts on this path (p_4).

Leacock and Chodorow's similarity[7]. It corresponds to a transformation of the Rada distance into a similarity. The shortest path between two concepts of the ontology restricted to taxonomic links is normalized by introducing a division by the double of the maximum hierarchy depth:

$$sim_{lc}(c_i, c_j) = -\log \frac{\min\limits_{p \in pths(c_i, c_j)} len_n(p)}{2 * \max\limits_{c \in cpts} \Big(\max\limits_{p \in pths(c, rt)} len_n(p) \Big)} \tag{3}$$

Like the Rada's measure, only the shortest path length (p_1) influences this measure.

Wu and Palmer's similarity[8]. It is a measure between concepts in an ontology restricted to taxonomic links. The two parameters which are the length of the two paths from c_i to $mscs(c_i, c_j)$ and from c_j to $mscs(c_i, c_j)$ in the Wu and Palmer's definition have been added. Their addition corresponds to the shortest path between c_i and c_j in the formula below:

$$sim_{wp}(c_i, c_j) = \frac{2 * \min\limits_{p \in pths(mscs(c_i, c_j), rt)} len_e(p)}{\min\limits_{p \in pths(c_i, c_j)} len_e(p) + 2 * \min\limits_{p \in pths(mscs(c_i, c_j), rt)} len_e(p)} \tag{4}$$

Here, the depth of the most specific common subsumer (p_2) has a non linear influence. We can observe this evolution if we set the shortest path length to a constant k : ($influence(x) = x/(x + k)$). Furthermore, this measure is sensitive to the shortest path (p_1).

Jiang and Conrath's distance[3]. The authors have used, like Resnik, a corpus in addition to the ontology restricted to taxonomic links. Jiang and Conrath formulate the distance between two concepts as the difference between the sum of the information content of the two concepts and the information content of their most specific common subsumer:

$$dist_{jc}(c_i, c_j) = 2 * \log P\Big(mscs(c_i, c_j)\Big) - \Big(\log P(c_i) + \log P(c_j)\Big) \tag{5}$$

This definition is composed of two interesting components which are the information content of the two concepts and the information content of their most specific common subsumer. We can suppose that it varies according to all the proposed parameters. But the combination revokes the effect of two parameters. Finally, this measure is sensitive to the shortest path length between c_i and c_j (p_1) and the density of concepts along this same path (p_3).

Lin's similarity[9]. Lin deduces from an axiomatic approach a measure based on an ontology restricted to taxonomic links and a corpus. This similarity takes into account the information shared by two concepts like Resnik, but also

the difference between them. The definition contains the same components as in the previous measure but the combination is not a difference but a ratio:

$$sim_l(c_i, c_j) = \frac{2 * \log P\Big(mscs(c_i, c_j)\Big)}{\Big(\log P(c_i) + \log P(c_j)\Big)} \tag{6}$$

In this case, the combination allows this measure to be sensitive to the whole parameter set (p_1, p_2, p_3, p_4). Lin notices that the Wu and Palmer measure is a particular case of his measure. Indeed if, for c' the father of c, we consider that $P(c|c')$ is constant, then we obtain the Wu and Palmer's measure.

Sussna's distance[2]. It is based on all the possible links. For each relation r, we define a weight $w(c_i \rightarrow_r c_j)$ from a given interval $[min_r; max_r]$. This weight is calculated with the local density which corresponds to the number of relations of the type r which go from c_i:

$$w(c_i \rightarrow_r c_j) = max_r - \frac{max_r - min_r}{n_r(c_i)} \tag{7}$$

Sussna defines the distance between two adjacent concepts. This link corresponds to the relations r and its inverse r'.

$$dist_s(c_i, c_j) = \frac{w(c_i \rightarrow_r c_j) + w(c_j \rightarrow_{r'} c_i)}{2 * \max\Big[\min_{p \in pths(c_i, rt)} len_e(p); \min_{p \in pths(c_j, rt)} len_e(p)\Big]} \tag{8}$$

This formula is defined for adjacent nodes only. To calculate the distance between two concepts, we have to sum the distances of all the links which compose the shortest path between these two concepts. The distance obtained is sensitive to three parameters: the shortest path length between c_i and c_j (p_1), the density of the concepts along this same path (p_2) and the shortest path length from the root to the most specific common subsumer of c_i and c_j (p_3).

Hirst and St Onge's relatedness[10]. It is based on an ontology. Hirst and St Onge distinguish four relation types between two concepts qualified of extra-strong, strong, medium and week. Hirst and St Onge propose a different way to calculate the relatedness functions of the relation type. In the following formula, C and K are two constants:

$$rel_{hs}(c_i, c_j) = \begin{cases} 3 * C(\text{extra-strong}); 2 * C(\text{strong}); 0(\text{week}); \\ C - \min_{p \in pths(c_i, c_j)} len_e(p) \\ \quad -K * \min_{p \in pths(c_i, c_j)} trn(p) \end{cases} (\text{medium}) \tag{9}$$

Synthesis and conclusion. The table 1 summarizes the studied characteristics and the influential parameters of the measures. The four parameters are independent and issued from the ontology. However, no proposed measure takes into account the whole parameter set without the use of a corpus. Concerning the introduction of a corpus in addition to an ontology when building a measure, we believe that this latter brings few information in comparison with its algorithmic complexity.

Table 1. Typology of semantic measures

	characteristics		Parameters			
	sources	semantics	p_1	p_2	p_3	p_4
Rada	ontology	distance	√			
Resnik	ontology+corpus	similarity		√		√
Leacock-Chodorow	ontology	similarity	√			
Jiang-Conrath	ontology+corpus	distance	√		√	
Wu-Palmer	ontology	similarity	√	√		
Lin	ontology+corpus	similarity	√	√	√	√
Sussna	ontology	distance	√	√	√	
Hirst-StOnge	ontology	relatedness	√			

In the next future, we plan to define a semantic similarity measure based on an ontology which will be sensitive to all parameters: the length and density of concepts of the shortest path between two concepts and of the shortest path from the root to their most specific common subsumer.

References

1. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5** (1993) 199–220
2. Sussna, M.: Word sense disambiguation for free-text indexing using a massive semantic network. In: Proceedings of the Second International Conference on Information and Knowledge Management. (1993) 67–74
3. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics. (1997)
4. Budanitsky, A.: Lexical semantic relatedness and its application in natural language processing. Technical report, Computer Systems Research Group - University of Toronto (1999)
5. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics **19** (1989) 17–30
6. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Volume 1. (1995) 448–453
7. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. In Fellbaum, C., ed.: WordNet: An electronic lexical database. MIT Press (1998) 265–283
8. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics. (1994) 133–138
9. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann (1998) 296–304
10. Hirst, G., StOnge, D.: Lexical chains as representation of context for the detection and correction of malaproprisms. In Fellbaum, C., ed.: WordNet: An electronic lexical database. MIT Press (1998) 305–332