A Review on Semantic Similarity

Montserrat Batet
Universitat Rovira i Virgili, Spain

David Sánchez
Universitat Rovira i Virgili, Spain

INTRODUCTION

The enormous development of the Information Society and the Word Wide Web has increased the interest of researchers in the automated understanding of electronic textual resources. The counter stone of textual understanding is the assessment of the *semantic similarity* between textual entities (e.g., words, sentences or documents). Semantic similarity aims at assessing a score that reflects the resemblance between the meanings of the compared entities, so that algorithms (e.g., classification, clustering, etc.) can seamlessly manage textual information from a numerical perspective. To do so, similarity measures exploit one or several information or knowledge sources and rely on different theoretical principles. Semantic similarity estimation has received a lot of attention in the last decade, also becoming a hot topic in many application areas such as natural language processing, information management and retrieval, textual data analysis and classification, and the Semantic Web.

Considering the plethora and heterogeneity of semantic similarity approaches available in the literature, this chapter offers researchers and practitioners aiming to develop or to exploit similarity measures: i) a description on the main notions involved in the assessment of semantic similarity, ii) a classification of the usual approaches proposed in the literature according to the theoretical principles on which they rely, iii) a critical comparison between these approaches, highlighting their advantages and shortcomings under the dimensions of accuracy, applicability and dependency on external resources, and iv) a list of research challenges.

BACKGROUND

According to the knowledge source used to extract semantic evidences to guide the similarity assessment, measures can be grouped in several families.

Ontology-based measures estimate the similarity of two concepts according to the structured knowledge offered by ontologies. They can be classified into:

1. *Edge-counting measures* evaluate the number of semantic links separating the two concepts in the ontology (Leacock & Chodorow, 1998; Li, *et al.*, 2003; Rada, *et al.*, 1989; Wu & Palmer, 1994). For example, Wu and Palmer compute similarity according to the number of taxonomic links (N_1 and N_2) between the two concepts (a, b) and their taxonomic ancestor, and the number of links (N_3) of that ancestor and the root node of the ontology, which acts as a normalization factor.

$$sim_{w\&p}(a,b)\frac{2\times N_3}{N_1+N_2+2\times N_3} \quad (1)$$

2. *Feature-based measures* rely on the amount of overlapping ontological features (e.g., taxonomic ancestors, concept descriptions, etc.) between the compared concepts (Petrakis, *et al.*, 2006; Rodríguez & Egenhofer, 2003; Sánchez, *et al.*, 2012a). For example, Sanchez and Batet measure the similarity between two concepts a and b according to the inverse non-linear ratio between their disjoint and total taxonomic ancestors $T(a)$ and $T(b)$.

DOI: 10.4018/978-1-4666-5888-2.ch746

$$sim_{s\&b}(a,b) = -\log\left(1 + \frac{\left|T(a) \cup T(b)\right| - \left|T(a) \cap T(b)\right|}{\left|T(a) \cup T(b)\right|}\right) \quad (2)$$

3. *Measures based on quantifying the amount of information* (i.e., Information Content (IC)) that concepts have in common (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995). Commonalties are extracted from the common taxonomic ancestors of the compared concepts, whereas the informativeness of concepts is computed either extrinsically from the concept occurrences in a corpus (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995) or intrinsically, according to the number of taxonomical descendants and/or ancestors modeled in the ontology (Sánchez & Batet, 2012; Seco, *et al.*, 2004). For example, Lin measures the similarity between concepts *a* and *b* according to the ratio between the informativeness of their *Least Common Subsumer* (*LCS*(*a*,*b*)) and the informativeness of each individual concept.

$$sim_{lin}(a,b) = \frac{2 \times IC(LCS(a,b))}{(IC(a) + IC(b))} \quad (3)$$

Distributional approaches only use textual corpora to infer semantics. They are based on the assumption that words with similar distributions have similar meanings (Waltinger, *et al.*, 2009). Thus, they assess term similarities according to their co-occurrence in corpora. As words may co-occur due to different kinds of relationships (i.e., taxonomic and non-taxonomic), distributional measures capture the more general notion of semantic *relatedness* in contrast to *similarity*, which is understood strictly as taxonomic resemblance. Distributional approaches can be classified into:

1. *First order co-occurrence measures* assume that related terms have a tendency to co-occur, and measure relatedness directly from their explicit co-occurrence (Bollegala, *et al.*, 2009; Cilibrasi & Vitányi, 2006; Turney, 2001). For example, Turney uses the Point-wise Mutual Information score (PMI) to measure the relatedness between terms according to the ratio between their probability of co-occurrence and the product of their individual probabilities.

$$PMI(a,b) = -\log\frac{p(a,b)}{p(a)p(b)} \quad (4)$$

2. *Second order co-occurrence measures* estimate relatedness as a function of the co-occurrence of words appearing in the context in which the compared terms occur (Banerjee & Pedersen, 2003; Patwardhan & Pedersen, 2006; Wan & Angryk, 2007). For example, Patwardhan and Pedersen use the WordNet glosses of the compared terms to construct their vectors of contexts (v_a, v_b) and measure their relatedness as the cosine of the angle between these vectors.

$$Relatedness(a,b) = \frac{\vec{v}_a \cdot \vec{v}_b}{\left|\vec{v}_a\right| \cdot \left|\vec{v}_b\right|} \quad (5)$$

DISCUSSION

In this section, the advantages and drawbacks of the different families of measures are pointed out, under the dimensions of expected accuracy, applicability and dependency on external resources. The discussion relies on the analytical, empirical and comparative results reported in the literature, which is thoughtfully cited, even though, due to space constraints, concrete results are not reproduced.

Edge-Counting Measures: Advantages and Limitations

In general, edge-counting measures are able to provide reasonably accurate results when a detailed and taxonomically homogenous ontology is used (Wu & Palmer, 1994). They have a low computational cost (compared to approaches relying on textual corpora) and they are easily implementable and applicable (Batet, *et al.*, 2011).

However, they just evaluate the shortest taxonomical path between concept pairs as evidences of distance (i.e., the opposite to similarity). This is a drawback, because many ontologies (e.g., WordNet, SNOMED-CT or MeSH) incorporate multiple taxonomical inheritance that would result in several taxonomical paths connecting concept pairs. Those paths represent explicit knowledge that is omitted by edge-counting methods (Batet, *et al.*, 2011). Due to their simplicity, they usually