

# Semantic Measures: How Similar? How Related?

Teresa Costa<sup>(✉)</sup> and José Paulo Leal

CRACS and INESC-Porto LA, Faculty of Sciences,  
University of Porto, Porto, Portugal  
{zp,teresa.costa}@dcc.fc.up.pt

**Abstract.** There are two main types of semantic measures (SM): similarity and relatedness. There are also two main types of datasets, those intended for similarity evaluations and those intended for relatedness. Although they are clearly distinct, they are similar enough to generate some misconceptions.

Is there a confusion between similarity and relatedness among the semantic measure community, both the designers of SMs and the creators of benchmarks? This is the question that the research presented in this paper tries to answer. Authors performed a survey of both the SMs and datasets and executed a cross evaluation of those measures and datasets. The results show different consistency of measures with datasets of the same type. This research enabled us to conclude not only that there is indeed some confusion but also to pinpoint the SMs and benchmarks less consistent with their intended type.

**Keywords:** Semantic similarity · Semantic relatedness · Semantic measures · Linked data

## 1 Introduction

Semantic measures are an attempt to quantify and compare pairs of concepts, words or sentences. They can be regarded as a kind of distance in a semantic space [1]. The object of semantic measures is inherently psychological, making an objective analysis more difficult. To complicate matters, there are two main kinds of semantic measures: similarity and relatedness. Similarity measures the amount of common features and relatedness ponders other kinds of relationships. Although these two kinds of semantic measures are distinct, are they defined and benchmarked in acceptable terms, so that they effectively measure different things?

Similarity and relatedness are indeed distinct concepts. The similarity of two concepts depends on size of the smallest class that contains them. Relatedness depends on any relationships connecting the two concepts, including but not restricted to class membership and inclusion. For instance, the concepts of dog and cat are similar insofar they are both mammals; the same can be said about ant and flea since they are both insects. An ant and a dog are similar insofar as they are both animals, but less similar than cats and dogs. This is so since

the class of animals contains both the classes of mammals and insects. Fleas are related to cats and dogs since they parasitize them, thus fleas are more related to dogs than ants. This is not because of the features they share, and they do share some since they are all animals, but because of other relationships, in this case parasitism. Thus, the similarity of dogs and fleas may be the same as the similarity of dogs and ants, but the relatedness of dogs and fleas is greater than that of dogs and ants. There is a clear difference between similarity and relatedness but people often confuse the two, or they value them in different ways. Based on a classical example [2], one could argue that some people value more similarity than relatedness.

There is growing evidence that the confusion between similarity and relatedness exists also among the researchers of semantic measures [1]. There are cases of semantic measures that are designed for similarity and then validated using relatedness datasets benchmark [3–5]. Arguably the source of this confusion is the perception that similarity is a particular case of relatedness [1] (page 15). In fact, similarity is based on *is-a* relationships and these are a particular kind of the relationships that may be considered in relatedness. However, this does *not* entail that a similarity measure is a particular case of a relatedness measure. As a metaphor, consider the routes available on a digital map between 2 given points  $a$  and  $b$  by different means of transport – walking, public transportation or car – and their respective times. These can be named  $t_w(a, b)$ ,  $t_p(a, b)$  and  $t_c(a, b)$ . One can add a fourth route – the quickest one, or  $t_q(a, b)$  – which can be obtained with a different means of transport according to each pair of points. Although the car is a particular means of transport, that in some cases is the quickest means of transport, that does not entail that  $t_c(a, b)$  is a particular case of  $t_q(a, b)$ .

By the same token a similarity measure using only *is-a* relationships is not a particular case of a relatedness measure considering all kinds of relationships, including the former. In particular, it does not make sense to use a relatedness dataset as benchmark for a similarity measure. The respondents of the questionnaires used to create a dataset received a clear set of instruction (we hope) stating what is similarity, what is relatedness, and how they differ. Thus a measure should not be compared with an estimation of a different type.

To better understand the tension between similarity and relatedness in semantic measures and benchmarks, the authors surveyed several path-based semantic measures and datasets, described on Sect. 2. Details of the implementation of these measures are provided on Sect. 3 and the results of the cross evaluation are described on Sect. 4. Section 5 summarizes the presented work, showing evidences of a misconception between similarity and relatedness.

## 2 Background

Semantic measures evaluate the strength of the semantic relationships between elements (words, concepts, phrases). This evaluation relies on the analysis of information extracted from semantic sources.

The type of the semantic measure depends on the type of semantic source. There are two kinds of semantic sources, the unstructured or semi-structured ones (plain texts and dictionaries, for instance) that are used by Distributional measures and the structured ones, that are used by Knowledge-based measures.

Knowledge-based measures rely on knowledge representations, namely semantic graphs. They estimate the semantic measures by taking advantage of the structural properties of the graph, comparing elements by studying their interconnections and the semantics carried in those relationships. These measures follow three different approaches: the structural approach (e.g. [6–8]), the featured-based approach (e.g. [9]) and the Information Theoretical approach (e.g. [10]).

Path-based measures follow the structural approach. They take advantage of several graph traversal strategies, such as shortest path, random walks or other interaction analysis. These measures focus on the analysis of the interconnections between nodes and use it to estimate the similarity (or relatedness) between them.

Several semantic similarity [5–8] and semantic relatedness [4, 11] measures were evaluated on this work. These measures rely on the definition of *shortest path* and *least common subsumer*.

The accuracy of a semantic measures is usually evaluated on how well it mimics the human capacity of comparing things. Datasets used in this validation process average human ratings for a set of words [1]. Those scores can be either of similarity or relatedness, as described on the instruction provided to the people that evaluated the dataset.

This work considered 4 semantic similarity datasets [12–14, 16] and 5 semantic relatedness datasets [14, 17–20].

### 3 Implementation

In the previous Section, several semantic measures were described. With the exception of the Hirst and St-Onge measure, they were originally designed to measure semantic similarity. However, those measures were adapted to estimate semantic relatedness, as proposed by Strube and Ponzetto [4].

In addition to these measures, Resnik similarity and Hirst and St-Onge relatedness were also adapted, the former to compute relatedness and the later to compute similarity, using an approach similar to that of Strube and Ponzetto. To compute relatedness using Resnik method one must use all the available properties instead only the taxonomic ones. To compute similarity using the Hirst and St-Onge method one must limit the shape of the allowable paths (to up and down), and also limit the properties in the upwards and downwards categories to the taxonomic ones.

All the described measures were implemented to compute both similarity and relatedness. The implementation process considered the following assumptions:

- the value of the semantic measure between a word and itself is its maximum value;

- the value of the semantic measure between two words, if one is not in the semantic proxy, is its minimum value;
- if the semantic proxy has no root or has several ones, a new node is inserted to form a semantic tree with a single root;
- the disambiguation strategy selects the pair of concepts (derived from the two input words) that produces the best measure.

All the semantic measures detailed on Sect. 2 depend on a graph traversal to search the best path connecting two different nodes. This can be a very time consuming process, in particular if a remote source is used. Knowledge bases, such as WordNet<sup>1</sup> [15], usually provide dumps of their data. These dumps were used to preprocess the semantic graph and store it locally. This task was performed using the RDF data dumps available for each version of WordNet.

A testbed to computed semantic measures was developed to support the validation process and is freely available online<sup>2</sup>. It is a Java Web Application created using the Google Web Toolkit, with a back-end server that stores the preprocessed graphs and computes the measures, and a front-end responsible for user interaction. The user interface allows the selection of semantic methods, semantic proxies, and a pair of words. After computation, the best result is displayed for each measure. This consists of the measure value, the pair of concepts associated to the given words, and the path linking them. If available, the user can browse other concept pairs with alternative values.

## 4 Validation

The cross validation process presented in this section used 10 different semantic measures (5 similarity and 5 relatedness) and 9 semantic datasets (4 similarity and 5 relatedness). As knowledge proxy, the three latest versions of WordNet were used.

The following tables summarize the results obtained for each WordNet version. Each measure as two variants, similarity and relatedness, respectively represented by an *S* and an *R* in the table row header. Datasets are also divided into similarity and relatedness. Thus rows are associated with measures and columns with datasets. The values on the cells are Spearman's rank order correlations between the computed values of the row's measure with the column's dataset values. The checkmark symbol (✓) means that the obtained result matches the expectations, which means that the semantic measure of a type performs better for a dataset of that type.

Table 1 presents the results obtained for the WordNet 2.1. WUP and HSO similarity measures stand out since they correctly identify the 4 similarity benchmarks. The other measures have mediocre results for datasets of the same type. The dataset with best performance is WS Sim that is correctly identified by all measures while MTurk-287 and MEN are always misidentified.

<sup>1</sup> [Wordnet.princeton.edu](http://wordnet.princeton.edu).

<sup>2</sup> <http://quilter.dcc.fc.up.pt/smcomp>.

**Table 1.** Cross evaluation of the semantic measures and semantic benchmarks using WordNet 2.1 as semantic source.

|        |   | Similarity datasets |        |        |            | Relatedness datasets |        |           |           |      |
|--------|---|---------------------|--------|--------|------------|----------------------|--------|-----------|-----------|------|
|        |   | MC30                | RG65   | WS Sim | SimLex 999 | WS353                | WS Rel | MTurk 287 | MTurk 771 | MEN  |
| Rada   | S | 0.21                | 0.28   | 0.41 ✓ | 0.20       | 0.22                 | 0.06   | 0.27      | 0.28      | 0.22 |
|        | R | 0.26                | 0.37   | 0.27   | 0.22       | 0.18                 | 0.05   | 0.22      | 0.33 ✓    | 0.15 |
| LCH    | S | 0.21                | 0.28   | 0.41 ✓ | 0.20       | 0.22                 | 0.06   | 0.27      | 0.28      | 0.22 |
|        | R | 0.26                | 0.37   | 0.26   | 0.22       | 0.18                 | 0.05   | 0.22      | 0.33 ✓    | 0.15 |
| WUP    | S | 0.20 ✓              | 0.27 ✓ | 0.39 ✓ | 0.10 ✓     | 0.20                 | 0.03   | 0.26      | 0.18      | 0.18 |
|        | R | 0.14                | 0.12   | 0.17   | -0.08      | 0.12                 | 0.10 ✓ | 0.17      | 0.10      | 0.07 |
| Resnik | S | 0.2                 | 0.25   | 0.35 ✓ | 0.25 ✓     | 0.19                 | 0.04   | 0.25      | 0.16      | 0.17 |
|        | R | 0.26                | 0.37   | 0.26   | 0.22       | 0.18                 | 0.05 ✓ | 0.22      | 0.33 ✓    | 0.15 |
| HSO    | S | 0.22 ✓              | 0.43 ✓ | 0.40 ✓ | 0.24 ✓     | 0.23                 | 0.05   | 0.31      | 0.29      | 0.29 |
|        | R | 0.18                | 0.27   | 0.34   | 0.14       | 0.23 ✓               | 0.11 ✓ | 0.29      | 0.18      | 0.25 |

**Table 2.** Cross evaluation using WordNet 3.0 as semantic source.

|        |   | Similarity Datasets |        |        |            | Relatedness Datasets |        |           |           |      |
|--------|---|---------------------|--------|--------|------------|----------------------|--------|-----------|-----------|------|
|        |   | MC30                | RG65   | WS Sim | SimLex 999 | WS353                | WS Rel | MTurk 287 | MTurk 771 | MEN  |
| Rada   | S | 0.15                | 0.26 ✓ | 0.34 ✓ | 0.20       | 0.20                 | 0.04   | 0.26      | 0.27      | 0.22 |
|        | R | 0.29                | 0.23   | 0.28   | 0.23       | 0.20 ✓               | 0.11 ✓ | 0.14      | 0.33 ✓    | 0.16 |
| LCH    | S | 0.16                | 0.26 ✓ | 0.37 ✓ | 0.20       | 0.20                 | 0.05   | 0.26      | 0.27      | 0.22 |
|        | R | 0.29                | 0.23   | 0.28   | 0.23       | 0.20 ✓               | 0.11 ✓ | 0.14      | 0.33 ✓    | 0.16 |
| WUP    | S | 0.13 ✓              | 0.23 ✓ | 0.33 ✓ | 0.08 ✓     | 0.18                 | 0.04   | 0.24      | 0.16      | 0.18 |
|        | R | 0.12                | 0.19   | 0.13   | -0.09      | 0.06                 | 0.01   | 0.11      | 0.09      | 0.08 |
| Resnik | S | 0.15                | 0.21   | 0.31 ✓ | 0.14       | 0.16                 | 0.02   | 0.25      | 0.15      | 0.18 |
|        | R | 0.29                | 0.23   | 0.28   | 0.23       | 0.20 ✓               | 0.11 ✓ | 0.14      | 0.33 ✓    | 0.16 |
| HSO    | S | 0.16 ✓              | 0.41 ✓ | 0.39 ✓ | 0.24 ✓     | 0.23                 | 0.06   | 0.26      | 0.28      | 0.3  |
|        | R | 0.11                | 0.26   | 0.32   | 0.14       | 0.21                 | 0.09 ✓ | 0.26 ✓    | 0.18      | 0.25 |

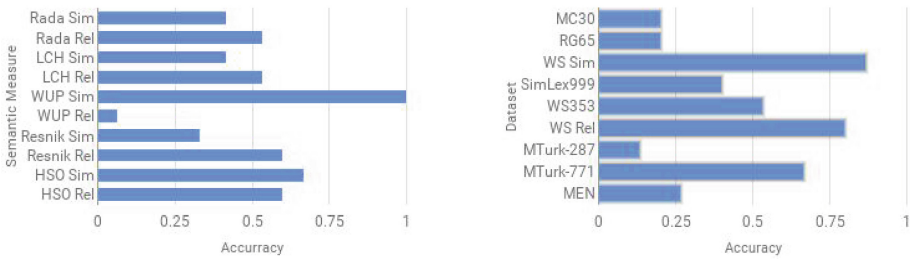
Table 2 presents the results obtained for the WordNet 3.0. WUP and HSO similarity measures stand out again since they identify correctly the 4 similarity benchmarks. The other measures have average results for datasets of the same type. The dataset with best performance is WS Sim that is correctly identified by all measures while MEN is always misidentified.

Table 3 presents the results obtained for the WordNet 2.1. WUP similarity measure stands out since it correctly identify the 4 similarity benchmarks. HSO relatedness measure also stands out by identifying all the relatedness datasets. The other measures have mediocre results for datasets of the same type. The dataset with best performance is WS Rel that is correctly identified by all measures. All benchmarks have their types correctly identified at least once.

The bar graphs of Fig. 1 provide an overview of the accuracy of semantic measures and datasets across the 3 WordNet versions. From the semantic measures perspective, the WUP measures has the best and worst results in similarity and relatedness respectively. It should be noted that the original measure was designed for similarity. All the other measures have mediocre results, with around

**Table 3.** Cross evaluation using WordNet 3.1 as semantic source.

|        |   | Similarity datasets |        |        |            | Relatedness datasets |        |           |           |        |
|--------|---|---------------------|--------|--------|------------|----------------------|--------|-----------|-----------|--------|
|        |   | MC30                | RG65   | WS Sim | SimLex 999 | WS353                | WS Rel | MTurk 287 | MTurk 771 | MEN    |
| Rada   | S | 0.6 ✓               | 0.76   | 0.57 ✓ | 0.37       | 0.28                 | -0.04  | 0.3       | 0.42      | 0.3    |
|        | R | 0.48                | 0.78   | 0.46   | 0.46       | 0.32 ✓               | 0.13 ✓ | 0.25      | 0.42 ✓    | 0.45 ✓ |
| LCH    | S | 0.6 ✓               | 0.76   | 0.57 ✓ | 0.37       | 0.28                 | -0.04  | 0.3       | 0.42      | 0.3    |
|        | R | 0.48                | 0.79   | 0.45   | 0.47       | 0.33 ✓               | 0.12 ✓ | 0.25      | 0.42 ✓    | 0.44 ✓ |
| WUP    | S | 0.66 ✓              | 0.7 ✓  | 0.52 ✓ | 0.25 ✓     | 0.27                 | -0.03  | 0.28      | 0.27      | 0.25   |
|        | R | 0.58                | 0.65   | 0.31   | 0.12       | 0.20                 | 0.08 ✓ | 0.21      | 0.16      | 0.21   |
| Resnik | S | 0.59                | 0.73 ✓ | 0.46   | 0.3        | 0.21                 | -0.08  | 0.29      | 0.32      | 0.24   |
|        | R | 0.59                | 0.70   | 0.70   | 0.41       | 0.25 ✓               | 0.16 ✓ | 0.24      | 0.35 ✓    | 0.34 ✓ |
| HSO    | S | 0.38                | 0.52   | 0.45   | 0.33       | 0.23                 | -0.03  | 0.32      | 0.27      | 0.33   |
|        | R | 0.62                | 0.65   | 0.48   | 0.43       | 0.31 ✓               | 0.11 ✓ | 0.34 ✓    | 0.27 ✓    | 0.38 ✓ |



**Fig. 1.** Datasets accuracy

50 % of accuracy rate. From the dataset perspective, two datasets stand out from the pack with accuracy rate above 75 %: the twin datasets WS Sim and WS Rel.

These results show that there may be some misconception regarding similarity and relatedness among the semantic measure community, both on the measure designers and on the data set creators. However, there are measures and benchmarks that stand out for their accuracy.

## 5 Conclusions

Semantic measures quantify the relationship between concepts, words and sentences. They try to mimic the human capacity for comparing things, hindering the analysis of artificial SM. There are semantic measures that estimate the amount of features two elements share – similarity – or that estimate all type of relationships between them – relatedness.

Despite being two different concepts, there seems to exist some confusion between them, namely among the semantic measures community. There are cases of semantic datasets that are wrongly categorized and cases of semantic measures that are designed for similarity, but evaluated using semantic relatedness datasets.

This paper surveyed several well known semantic benchmarks and path-based measures. Aiming to understand the tension between similarity and relatedness, a cross evaluation was performed using all measures (and their adaptations) with all surveyed datasets. This process was executed with three different versions of WordNet as semantic proxy. Assuming that there is no confusion between similarity and relatedness, it should be possible to use semantic measures of both types to identify the type of a semantic dataset. It should be also possible to use semantic benchmarks of the two different types to categorize a semantic measure.

The validation showed that this is not the case. In fact, the opposite is more frequent. Most of the SMs do not guess correctly the datasets of their types and vice-versa. This enables us to conclude that some misconception regarding relatedness and similarity may exist among the semantic measure community. Fortunately, this research allowed us to pinpoint a few cases where SMs and datasets are more accurate, namely the WUP similarity measure and the WS-Sim and WS-Rel datasets.

**Acknowledgments.** This work is partially financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by the FCT within project POCI-01-0145-FEDER-006961 and project “NORTE-01-0145-FEDER-000020” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the European Regional Development Fund (ERDF).

## References

1. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis. *Synth. Lect. Hum. Lang. Technol.* **8**, 1–254 (2015)
2. Gorodnichenko, Y., Roland, G.: Understanding the individualism-collectivism cleavage, its effects: lessons from cultural psychology. *Institutions Comp. Econ. Dev.* **150**, 213 (2012)
3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.* **32**, 13–47 (2006)
4. Strube, M., Ponzetto, S.: WikiRelate! Computing semantic relatedness using wikipedia. In: *AAAI* (2006)
5. Philip, R.: Using information content to evaluate semantic similarity in a taxonomy. In: *IJCAI* (1995)
6. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.* **19**, 17–30 (1989)
7. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. *WordNet: Electr. Lexical Database* **49**, 265–283 (1998)
8. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (1994)
9. Bodenreider, O., Aubry, M., Burgun, A.: Non-lexical approaches to identifying associative relations in the gene ontology. In: *Pacific Symposium on Biocomputing* (2005)

10. Lin, D.: An information-theoretic definition of similarity. In: ICML (1998)
11. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: Electr. Lexical Database* **305**, 305–332 (1998)
12. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. *Commun. ACM* **8**, 627–633 (1965)
13. Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. *Lang. Cogn. Proc.* **6**, 1–28 (1991)
14. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A.: A study on similarity, relatedness using distributional, wordnet-based approaches. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2009)
15. Fellbaum, C.: *WordNet*. Wiley, New York (1999)
16. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: evaluating semantic models with (genuine) similarity estimation (2014). arXiv preprint [arXiv:1408.3456](https://arxiv.org/abs/1408.3456)
17. Evgeniy, G.: The WordSimilarity-353 Test Collection. <http://www.cs.technion.ac.il/gabr/resources/data/wordsim353/>
18. Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S.: A word at a time, computing word relatedness using temporal semantic analysis. In: *Proceedings of the 20th International Conference on World Wide Web* (2011)
19. Halawi, G., Dror, G., Gabrilovich, E., Koren, Y.: Large-scale learning of word relatedness with constraints. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012)
20. Bruni, E., Tran, N.-K., Baroni, M.: Multimodal distributional semantics. *J. Artif. Intell. Res. (JAIR)* **49**, 1–47 (2014)