Natural Language Engineering **24** (6): 861–886. © Cambridge University Press 2018
doi:10.1017/S1351324918000190

*Weighting-based semantic similarity measure based on topological parameters in semantic taxonomy**

ABDULGABBAR SAIF[1], UMMI ZAKIAH ZAINODIN[2], NAZLIA OMAR[2] and ABDULLAH SAEED GHAREB[1]

[1]*Department of Computer Information Systems, Faculty of IT&CS, University of Saba Region, Marib, Yemen*
e-mail: agmssaif@gmail.com, aghurieb@usr.ac
[2]*Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Malaysia*
e-mail: ummizakiahzainodin@gmail.com, nazlia@ukm.edu.my

(*Received 25 July 2017; revised 30 March 2018; accepted 9 April 2018; first published online 4 June 2018*)

Abstract

Semantic measures are used in handling different issues in several research areas, such as artificial intelligence, natural language processing, knowledge engineering, bioinformatics, and information retrieval. Hierarchical feature-based semantic measures have been proposed to estimate the semantic similarity between two concepts/words depending on the features extracted from a semantic taxonomy (hierarchy) of a given lexical source. The central issue in these measures is the constant weighting assumption that all elements in the semantic representation of the concept possess the same relevance. In this paper, a new weighting-based semantic similarity measure is proposed to address the issues in hierarchical feature-based measures. Four mechanisms are introduced to weigh the degree of relevance of features in the semantic representation of a concept by using topological parameters (edge, depth, descendants, and density) in a semantic taxonomy. With the semantic taxonomy of WordNet, the proposed semantic measure is evaluated for word semantic similarity in four gold-standard datasets. Experimental results show that the proposed measure outperforms hierarchical feature-based semantic measures in all the datasets. Comparison results also imply that the proposed measure is more effective than information-content measures in measuring semantic similarity.

1 Introduction

Semantic similarity measurement has received considerable attention from researchers in cognitive science and computational semantic analysis. Researchers focus on different techniques on how humans estimate the semantic similarity between

*This work was partially funded by the Ministry of Higher Education in Malaysia under the grant no. (FRGS/1/2016/ICT02/UKM/02/11). The first author would like to thank the University of Saba Region for its financial supports.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

two mental representations (objects) in a psychological space (Griffiths, Steyvers and Tenenbaum 2007). The three main models used to measure the semantic similarity between two objects are spatial, structure based, and feature based. The main idea behind the spatial model is that objects are represented as points in a multidimensional metric space. The structure-based model, also called the network model, depends on the representation of concepts as nodes linked by edges that indicate pairwise relationships. The feature-based model is based on the assumption that a concept is represented as a set of features that reflect its characteristics, properties, or aspects.

Several studies have used feature-based methods to measure semantic similarity. The central role of feature-based methods in determining semantic similarity is based on the features selected to represent the semantics of concepts and the form of metrics used to measure the similarity between two representative features. In Cross, Yu and Hu (2013), Jiang *et al.* (2015), and Saif, Ab Aziz and Omar (2017), the experimental results showed that the performance of semantic similarity measures crucially depends on the features selected to represent each concept. In lexical sources, semantic features can be classified into two main types: local and hierarchical features. Local features that have been exploited in previous studies include a set of synonyms, direct lexical relations (hypernyms, hyponyms, meronyms, and holonyms), and the words in its gloss (textual definition of the concept) in the WordNet (WN) lexicon, redirect pages, the first paragraph of an article, hyperlinks, and categories as representative features of a given article in Wikipedia (Saif, Ab Aziz and Omar 2014; Jiang *et al.* 2015; Saif *et al.* 2017). Hierarchical features require representation of the lexical source as a semantic taxonomy, in which all concepts in the lexical sources are connected either directly or indirectly.

Semantic measures based on hierarchical features (Gentleman 2005; Gurevych 2005; Batet, Sánchez and Valls 2011; Saif *et al.* 2014; Saif *et al.* 2017) (hereafter referred to as hierarchical feature-based semantic measures) merge structure-based and feature-based methods to estimate the similarity between two concepts. Thus, the structure-based method is used to represent knowledge in a given lexical source as a taxonomy using 'is–a' relationships. Then, each concept is semantically represented by a set of features recursively extracted from a taxonomy. Subsequently, semantic similarity is measured based on shared and non-shared features. The experimental results (Saif *et al.* 2014; Saif *et al.* 2017) showed that hierarchical feature-based semantic measures outperform measures based on local features. However, the main limitation of hierarchical feature-based semantic measures is that it implicitly assumes that every element in the representative features of a certain concept has the same relevance as the concept. Although the features of these measures are hierarchically extracted to represent the meaning of a concept, the topological (structural) parameters, such as the depth and number of descendants, that reflect the specificity or generality of the concepts are ignored. These topological parameters are important in measuring semantic similarity and have been widely used to address specificity in path-based and information content (IC)-based measures (Lastra-Díaz and García-Serrano 2015b). Consequently, ignoring these topological parameters in the hierarchical feature-based measure may affect the performance of this measure.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

In this study, a new weighting-based semantic similarity measure is proposed to assess the semantic similarity between two concepts/words by using knowledge encoding in lexical sources to overcome the shortcomings of hierarchical feature-based semantic measures. From the semantic taxonomy of a lexical source, the semantics of a concept is represented as a set of features extracted with hypernym relations. In the proposed measure, four weighting mechanisms are employed; these mechanisms use taxonomical parameters (edge, depth, density, and descendants) to score the degree of a feature's importance in the semantic representation of a certain concept. The proposed measure is evaluated with the semantic taxonomy of extensive knowledge sources exploited for semantic analysis (WN). In addition, the proposed measure, hierarchical feature-based measures, and feature-based measures are compared.

2 Related work

Feature-based measures refer to methods that represent a concept or synset by using semantic features encoded in structured knowledge sources. The measurement of similarity depends on set-based theory to determine the amount of shared and non-shared features between two representations of concepts. The first direction of feature-based measures represents a concept as a bag of words using the lexical semantic content in lexical sources. The earliest method (Lesk 1986) in this direction is a technique that employs glosses in lexical sources. Several studies have been conducted to improve this method by using other semantic features. For example, Banerjee and Pedersen (2003) used extended glosses by enriching features from the glosses of related concepts (hypernym, hyponym, meronym, holonym, and troponym). Gurevych (2005) represented concepts by using features that are closely related (synonymy, hypernyms, and meronyms) to the described concept to overcome the limitation of incomplete lexical sources, such as GermaNet. Using the same idea for the Arabic WordNet (AWN), Saif *et al.* (2014) used all direct relations to extract the features that represent the semantics of a concept. Saif *et al.* (2017) extended the semantic representation of a concept by using the semantic taxonomy of AWN on the basis of the hierarchical relations among synsets. For each node (concept), representation was generated by concatenating lemmas in its related concepts (ancestors, children, and siblings) as the textual representation (sequence of words). Jiang *et al.* (2015) proposed a feature-based semantic measure using structures of Wikipedia as semantic features. In their work, they defined a concept (Wikipedia article) using four features, namely, the title of the article, the redirect pages, the first paragraph of the article, and the internal hyperlinks. Then, the semantic similarity between two concepts was computed as the combination of the semantic similarities of these features using twelve techniques. All techniques depended on set-based theory to determine similarity using the intersection and union of two sets.

The second direction of feature-based measures depends on extracting features from the semantic taxonomy of lexical sources as a bag of concepts instead of as a bag of words in the first direction. For example, Gentleman (2005) introduced

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

a hierarchical feature-based semantic measure based on the shared knowledge of ancestors between two concepts extracted from a semantic taxonomy. This measure is defined as follows:

$$\mathrm{Sim}_{\mathrm{Gen}}(c_1,\ c_2) = \frac{|\mathrm{ancestor}(c_1) \cap \mathrm{ancestor}(c_2)|}{|\mathrm{ancestor}(c_1) \cup \mathrm{ancestor}(c_2)|} \tag{1}$$

where ancestor(c) refers to nodes in the paths from node c that corresponds to a given concept to the *root* of a semantic taxonomy. Batet *et al.* (2011) introduced a similar hierarchical feature-based semantic measure that depends on the amount of common information from the ancestors of concepts in a semantic taxonomy. In this measure, a concept is represented as a set of super-concepts that include the ancestors of a concept as well as the concept itself in semantic graph C, as follows:

$$T(c_i) = \{c_j \in C : c_j \text{ is an ancestor of } c_i\} \cup \{c_i\} \tag{2}$$

Therefore, the semantic similarity between two concepts c_1 and c_2 is defined as the ratio between the amount of non-shared concepts and the sum of shared and non-shared concepts as follows:

$$\mathrm{Sim}_{\mathrm{Batet}}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \tag{3}$$

For knowledge-based semantic analysis, several text-based methods have been introduced to weigh the features in semantic representation. In these methods, the semantics of a word is expressed by explicit concepts (features) in language resources as a weighted vector. In Gabrilovich and Markovitch (2007), the explicit semantic analysis (ESA) method was proposed for weighting features using the term frequency with inverse document frequency (*TF-IDF*) function. Many studies have modified the ESA method to improve the weights of concepts in semantic representation. For example, Hassan and Mihalcea (2009) utilized a Wikipedia category graph to boost the weight of concepts in the original ESA representation vector. The idea behind this modification is that specific concepts are more important than general concepts. Given that the categories can be organized hierarchically as a taxonomy based on subcategory links, the articles are inserted as leaf nodes to their corresponding categories, in which the nodes at the top level of the taxonomy are considered more general than the nodes at the bottom level. Thus, the weights of concepts in the representation vector are scaled according to their maximum depth (distance to the root) in the taxonomy. Radinsky *et al.* (2011) introduced the temporal semantic analysis method that integrates the distributions of the words and the *temporal dynamics* (concept's time series) of the concepts from an external background information resource to weigh the concept. Each concept's time series is scaled according to its original weight in the ESA vector. By using hyperlinks in Wikipedia, a salient semantic analysis (Hassan and Mihalcea 2011) was proposed to define the weight of the features in the semantic representation of a word by the point-wise mutual information of the associated salient concept and the word. The main idea behind this weighting is inspired by the distributional hypothesis 'you shall know a word by the company it keeps' (Firth 1957) that has been used in distribution-based approaches.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Taieb, Ben Aouicha and Ben Hamadou (2013) proposed the category semantic depiction method for representing the semantics of words over Wikipedia categories instead of articles in the ESA method. Each Wikipedia category is represented as a vector containing the terms that occur in the articles, which are assigned to a certain category. The terms of a certain article are defined as the following set: $T(a) = \{f_1, f_2, \ldots, f_m\}$, where f_i indicates the frequency of the ith term in article a and m is the number of terms in article a. The frequency of each term in $T(c)$ is the weight of the term in the semantic representation of the category that is calculated as the sum of the frequency of the term in all articles that belong to the category. The reduced semantic representation method (Saif, Ab Aziz and Omar 2016) has been proposed to handle the high dimensionality in the ESA method. The semantic representation of a word was defined as the concepts (features) containing this word in their definitions. In weighting and feature reduction, the semantic interpretation of words was constructed as vectors over the latent topics from the original ESA representation vectors. In modeling the latent topics, the latent Dirichlet allocation is adopted for the ESA vectors to extract the topics as probability distributions over the concepts rather than words in the traditional model.

Taxonomical parameters have been used in path-based measures (Al-Mubaid and Nguyen 2006; Wu *et al.* 2006; Wang and Hirst 2011) to handle the specificity issue and in intrinsic IC-based measures (Seco, Veale and Hayes 2004; Zhou, Wang and Gu 2008; Pirró 2009; Sánchez, Batet and Isern 2011; Taieb *et al.* 2011; Meng, Gu and Zhou 2012; Sánchez and Batet 2012; Yuan, Yu and Wang 2013; Harispe *et al.* 2014; Lastra-Díaz and García-Serrano 2015a; Lastra-Díaz and García-Serrano 2015b; Jiang *et al.* 2016) to address the issue of dependency on external knowledge sources (tagged corpora) for modeling the IC of concepts. The experimental results showed that these taxonomical parameters are useful in handling issues on measuring semantic similarity. Therefore, in this study, taxonomical parameters are used to handle the issue of the constant weighting assumption in hierarchical feature-based measures.

3 Weighted semantic measure

This section describes in detail the new proposed method for measuring semantic similarity between two words or concepts. With a lexical knowledge source, the proposed measure consists of three main steps, namely, semantic representation, weighting features, and measuring semantic similarity.

3.1 Semantic representation

Semantic representation considers how the meaning of the specific synset in a lexical source can be represented using semantic features. The main features in WN are lexical relations and glosses (textual definition of concepts). The most lexical relation that has been used in measuring semantic similarity is the 'is–a' relation (hypernyms/hyponyms). In the semantic taxonomy using the 'is–a' hierarchical relation, a synset inherits several characteristics from the super-concept and adds its

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

own specific features to represent its semantics. Thus, a synset is an accumulation of the propagated information from a distant ancestor to another less distant ancestor by adding specificities to each descendant (Aouicha, Taieb and Ezzeddine 2016). Subsequently, a synset depends strongly on its direct hypernyms and ancestors to form its meaning.

In this research, a synset is semantically represented using the super-concepts generated from hierarchical relations in a given lexical knowledge source. The motivation behind selecting this type of feature is that state-of-the-art experiments (Zhang, Gentile and Ciravegna 2013; Saif *et al.* 2017) have shown that this feature achieves promising results compared with other features. This finding is attributed to the fact that the definition of most concepts can be extracted from the taxonomical relations (hypernyms and hyponyms) of the described concept. For example, the definition of the concept *atropine* is introduced in WN as 'a poisonous crystalline alkaloid extracted from the nightshade family; used as an antispasmodic and to dilate the eye pupil; also administered in large amounts as an antidote for organophosphate nerve agents or organophosphate insecticides.' Thus, the words 'alkaloid,' 'antispasmodic,' 'antidote,' and 'agent' can be extracted from the ancestors of this concept. Based on the hyponyms, the terms 'domestic cats' and 'wildcats' can be extracted from the direct hyponyms of the concept *cat* that has the following gloss in WN: 'feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats.'

For each node (synset), its semantics is represented as a set of concepts that are extracted from its super-concepts in a semantic taxonomy. The super-concepts of a synset are extracted as nodes in the paths from the node that corresponds to this synset to the root of the semantic taxonomy as well as the synset itself. The reason for including the synset in its super-concept is to overcome the indistinguishability limitation. In this limitation, any two pairs of synsets having the same super-concepts will receive the same semantic similarity, which is not necessarily appropriate. For example, in Figure 1, the similarity between *antispasmodic*#1 and *remedy*#2 is equal to the similarity between the concept *remedy*#2 and itself. Formally, the form of the super-concepts of the synset in the semantic taxonomy T is defined as follows:

$$\text{Super}(syn) = \{c \in T \mid c \in \text{ancestors}(syn)\} \cup \{syn\} \tag{4}$$

where ancestors(syn) is a function that returns the hypernyms of syn in a recursive manner.

Given that a node can have more than one path to the root (multiple hypernym relations), the ancestors are extracted from each path. Therefore, as shown in Figure 1, the ancestors of the synset *atropine*#1 are *poison*#1, *antidote*#1, *alkaloid*#1, *mydriatic*#1, *antispasmodic*#1, *remedy*#2, *medicine*#2, *drug*#1, *organic compound*#2, *compound*#1, *agent*#2, *causal agent*#1, *substance*#1, and *entity*#1, which are extracted from the paths of the node *atropine*#1 to the root in the semantic taxonomy.

3.2 Weighting features

Based on the representation step, the features in the semantic representation of a specific synset are assumed to have the same degree of importance in the meaning

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

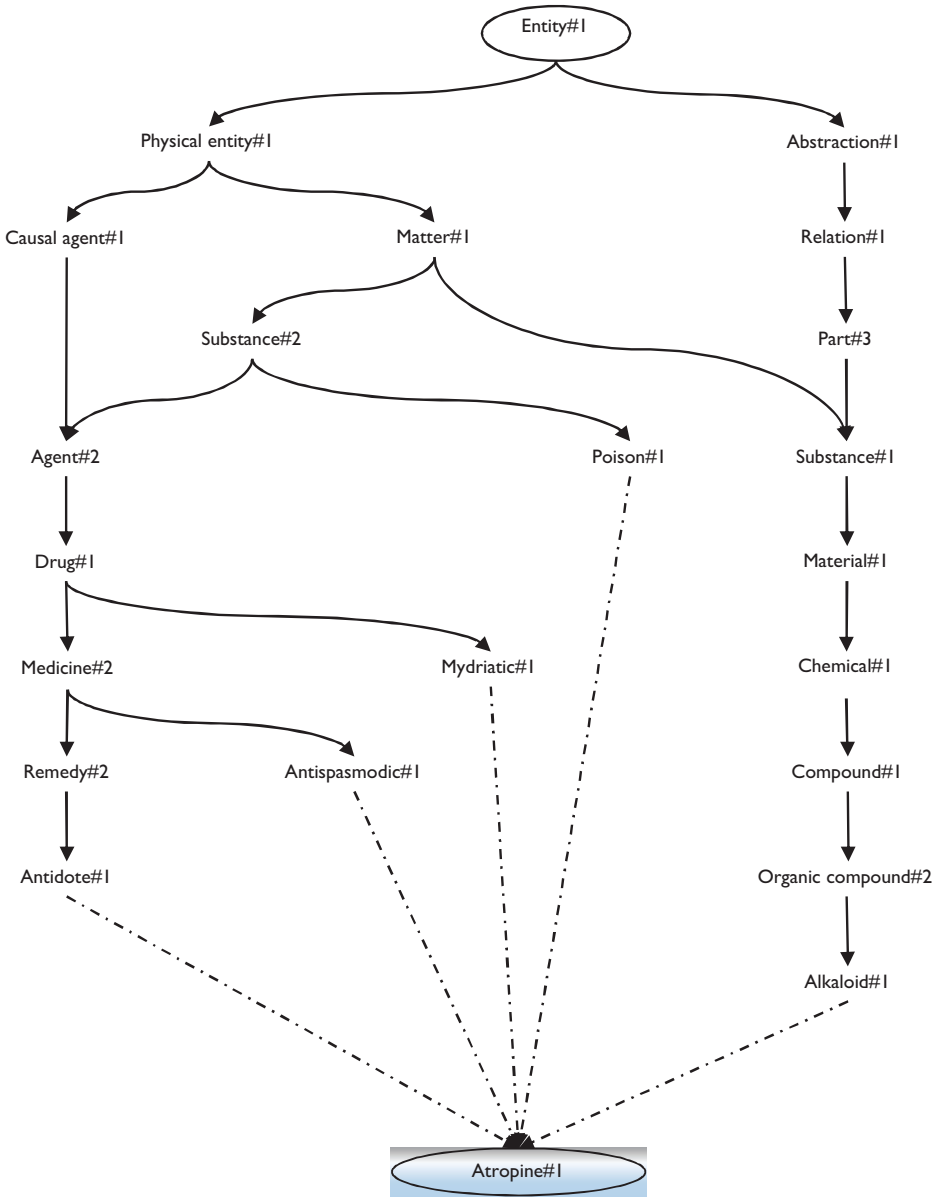


Fig. 1. (Colour online) Part of the semantic taxonomy of WN representing the super-concepts of *Atropine*#1.

of this synset. However, the semantic relations between a synset and its ancestors vary according to the distance between the synset and each concept of its ancestors; a nearer node is more semantically related to its nearest ancestors than a farther node. For instance, the concepts *Alkaloid*#1, *Mydriatic*#1, *Antispasmodic*#1, and *Antidote*#1 that correspond to the direct hypernyms of the synset *Atropine*#1 are more semantically related to the synset than *Entity*#1, *Substance*#1, and *Causal agent*#1, which have longer distance to the synset.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

In other words, the main limitation of this semantic representation is that it does not consider the specificity of concepts in a semantic hierarchy. In this study, the weighting step is introduced to overcome this limitation. In this step, each feature in the semantic representation of a given synset is weighted according to its importance to form the meaning of a synset. Formally, the semantic interpretation of a synset is represented by a row vector of concepts (its super-concepts) paired with their weights that reflect the degree of association of each concept to the synset. For a given synset *syn*, its semantic representation is defined based on the semantic taxonomy of the lexical knowledge source as follows:

$$R(syn) = \{(c_1, w_1), (c_2, w_2), \ldots, (c_m, w_m)\} \tag{5}$$

where $c_1, c_2, \ldots, c_m$ are the super-concepts of *syn* and $w_1, w_2, \ldots, w_m$ are the weights of the corresponding super-concepts.

In this step, four weighting mechanisms are introduced to quantify the semantic association between features in the semantic representation of a given concept and the concept itself. These mechanisms depend on the structural (topological) parameters of the semantic taxonomy of a given lexical source. The structural parameters include edge, depth, density, and descendants. These parameters have been used in many studies to address the specificity issue in path-based measures and to compute the IC in IC-based measures.

3.2.1 Edge

The first weighting mechanism depends on the distance between a given synset and each concept in its semantic representation. The distance metric has been exploited by many researchers to measure semantic similarity. In a semantic taxonomy, distance refers to the length of a path and is typically calculated by counting the number of edges along the path. Formally, the weight of concept *c* in the semantic representation of synset *syn* is defined as follows:

$$w(c) = \frac{1}{\text{length}\ (c,\ syn) + 1} \tag{6}$$

where the length returns the shortest path (number of edges) between *c* and *syn*.

3.2.2 Depth

The second weighting mechanism uses the depth of the concept in a semantic taxonomy as the weight of the features. The depth of synset *syn* denoted by depth(*syn*) is usually the number of nodes along the longest path between *syn* and root in the semantic taxonomy. Depth is significant in taxonomy-based semantic analysis because it indicates the degree of expressiveness/granularity of the hierarchy. The node at the top level is more general than the node at the bottom level. In other words, moving from a high level to another means moving from a narrow topic to the addition of specificities. The depth of concepts in WN ranges between 1 and 20. The histogram of depth in WN is illustrated in Figure 2.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 1. *Weights of the features of the semantic representation of the concept Atropine#1*

Features	Depth	*W* (depth)	Edges	*W* (edges)
Atropine#1	11	0.52	0.00	1.00
Alkaloid#1	10	0.50	1.00	0.50
Organic compound#1	9	0.48	2.00	0.33
Antidote#1	9	0.48	1.00	0.50
Remedy#2	8	0.45	2.00	0.33
Compound#1	8	0.45	3.00	0.25
Antispasmodic#1	8	0.45	1.00	0.50
Chemical#1	7	0.42	4.00	0.20
Mydriatic#1	7	0.42	1.00	0.50
Medicine#2	7	0.42	2.00	0.33
Drug#1	6	0.39	2.00	0.33
Material#1	6	0.39	5.00	0.16
Agent#2	5	0.35	3.00	0.25
Poison#1	5	0.35	1.00	0.50
Substance#1	5	0.35	6.00	0.14
Part#3	4	0.30	7.00	0.13
Substance#2	4	0.30	2.00	0.33
Matter#1	3	0.24	3.00	0.25
Causal agent#1	3	0.24	4.00	0.20
Relation#1	3	0.24	8.00	0.11
Physical entity#1	2	0.15	4.00	0.20
Abstraction#1	2	0.15	9.00	0.10
Entity#1	1	0.00	5.00	0.16

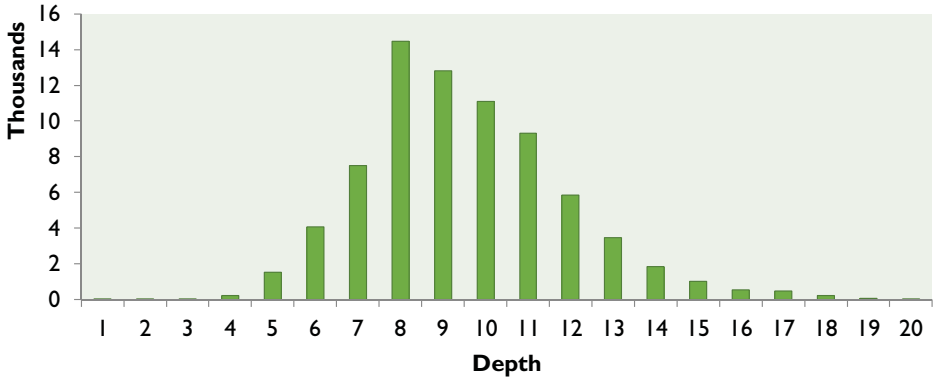


Fig. 2. (Colour online) Histogram of depth in the WN semantic taxonomy.

The difference between distance and depth is that the elements in the direct hypernyms (parents) of the specific node may have different depths in the semantic taxonomy. For example, in Figure 1, the node *atropine#1* has five direct hypernyms, namely, *alkaloid#1*, *antidote#1*, *antispasmodic#1*, *mydriatic#1*, and *poison#1*, which have depths of 10, 9, 8, 7, and 5, respectively. Meanwhile, the distance between *atropine#1* and each element in its hypernyms is 1. The weights of the features of the semantic representation of the concept *atropine#1* are presented in Table 1.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

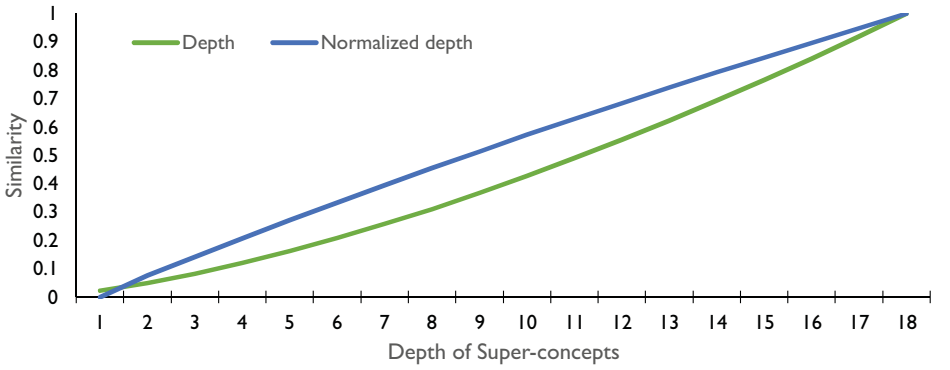


Fig. 3. (Colour online) Difference between depth and normalized depth in terms of linearity.

Although the weighting scores of the concepts in the super-concepts of the synset range linearly between 1 and the depth of this synset, the semantic similarity using these scores is not balanced across the concepts in the semantic representation. For this reason, the weighting scores are normalized to balance the semantic similarity scores across the different concepts in the semantic taxonomy. For the concepts in the representation, the normalization function for each concept is defined using the natural logarithm and the square root as follows:

$$w(c) = \log\left(\sqrt{\text{depth}(c)}\right) \tag{7}$$

Figure 3 shows the semantic similarity scores of the synset 'ox' and its super-concepts that have depths ranging from 1 to 18 using direct depth weighting and normalized weighting. In contrast to those of direct depth weighting, the similarity scores of using normalized weighting are linearly distributed.

3.2.3 *Descendants*

The third weighting mechanism exploits the number of descendants of a given concept and reflects the important parameters in semantic measures. The number of descendants of a concept is considered to be directly proportional to its degree of generality. A concept with a large number of descendants is considered more general than a concept with a small number of descendants. For example, in Figure 1, the concept 'physical entity' that has a descendant number of 18 is more general than the concept '*alkaloid*#1,' which has a smaller number of descendants (i.e., 1).

In weighting a concept, the *TF-IDF* technique that has been used in information retrieval is adopted in this work to quantify the strength of the relevance of concepts. In information retrieval, a document is represented as a vector, in which each dimension corresponds to the *TF-IDF* of the term. Inverse document frequency (IDF) is defined as $(\log \frac{n}{df})$, where n is the number of documents in the collection and df is the number of documents in the collection that contain the term. The term is considered general if it appears in many documents. Consequently, the weight of

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

concept c is defined as follows:

$$w(c) = \log\left(\frac{|T|}{|\text{descendant}(c)| + 1}\right) \tag{8}$$

where $|T|$ is the number of concepts in the semantic taxonomy or the number of descendants of the root of the semantic taxonomy as well as the root itself. Thus, the weight of the root in the semantic taxonomy is equivalent to zero.

3.2.4 Density

In a semantic taxonomy, the *local density* of a synset node refers to the number of its sibling nodes or the number of its child nodes (Wang and Hirst 2011; Zhang *et al.* 2013). The density parameter has been used in a few semantic similarity measures (Sussna 1993; Jiang and Conrath 1997) based on the assumption that two concepts are semantically close if they are densely connected locally. In other words, density is used as a parameter to determine the degree of specificity of a concept; the concepts in densely connected regions in the semantic taxonomy are considered highly specific and thus close to one another.

Based on the direct definition of local density, Wang and Hirst (2011) redefined the density of a concept to address information loss on the local density that reflects specificity. For example, the concept 'geological formation' has a density value of 37, whereas the concept 'shore,' which is one of its hyponyms, has a density value of 28 (less than the density of its parent). Intuitively, the specificity should monotonically increase as one descends the taxonomy. Thus, this study defines the accumulated density of a given concept to be inherited by its descendants as follows:

$$\text{density}_{\text{Accu}}(c) = \left(\sum_{c' \in \text{ancestors}(c)} \text{density}\left(c'\right)\right) + \text{density}(c) \tag{9}$$

where density(c) denotes the number of siblings of the concept as well as the concept itself. Figure 4 shows a comparison of the histograms of density and accumulated density in the semantic taxonomy of WN. The accumulated density is more normally distributed than the local density, given that the local density of more than 65 per cent of the nodes in the semantic taxonomy ranges between 1 and 6. Then, the features in the semantic representation of the given concept are weighted using the accumulated density as follows:

$$w(c) = \text{density}_{\text{Accu}}(c) \tag{10}$$

3.3 Semantic similarity

After the semantic signature representations for a pair of concepts have been obtained, the semantic similarity of two concepts is calculated by comparing their corresponding semantic representation vectors. In comparing the semantic representation vectors, the method that has been used extensively in previous work

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

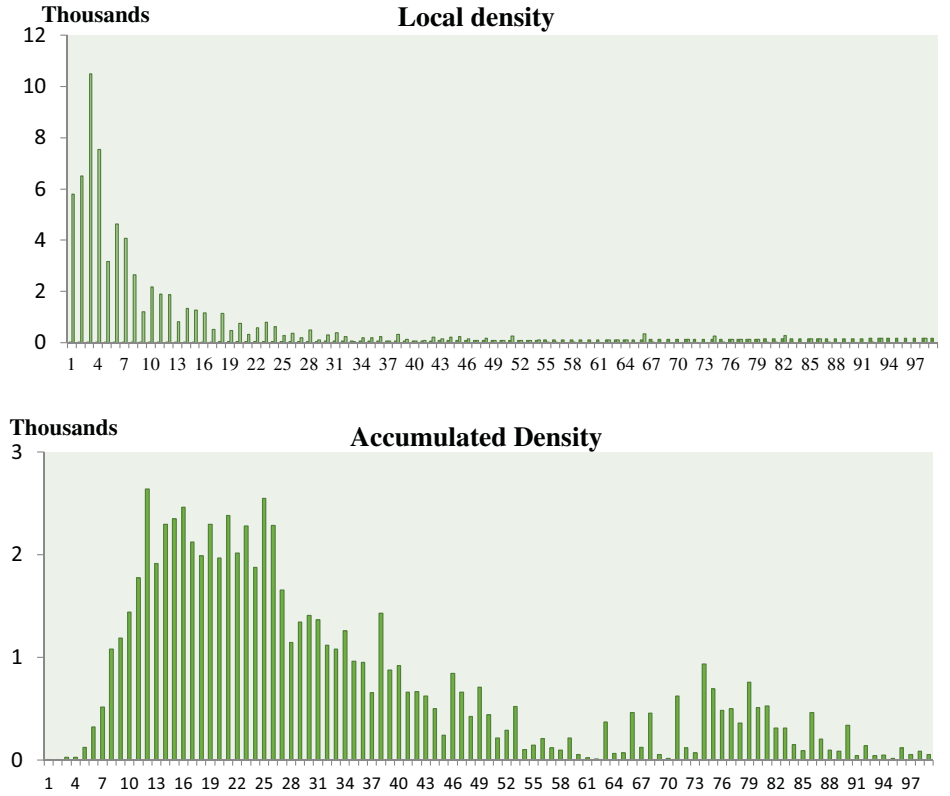


Fig. 4. (Colour online) Histograms of local density and accumulated density in WN.

to compare the vectors and the *cosine* is used in this research. The cosine metric is defined to measure the similarity between two concepts c_1 and c_2 as follows:

$$\mathrm{Sim}(c_1, c_2) = \frac{\sum_{c' \in R(c_1) \cap R(c_2)} w_1(c') \times w_2(c')}{\sqrt{\sum_{c' \in R(c_1)} w_1(c')^2} \sqrt{\sum_{c' \in R(c_2)} w_2(c')^2}} \tag{11}$$

where $w_1(c)$ and $w_2(c)$ are the weight of the concept c in semantic representation (R) of the concepts c_1 and c_2, respectively.

Given the ambiguity issue, polysemy words in WN are represented by several concepts (at least two concepts). Thus, the semantic similarity between two words depends on the similarity between their corresponding concepts. Following Sánchez *et al.* (2012) and Lastra-Díaz and García-Serrano (2015a), the semantic similarity between two words is computed in this work based on the maximum technique. First, all of the concepts that correspond to the senses of each word are extracted from WN. Second, the semantic similarity is computed for each pair of concepts in the combination of candidates. Finally, the maximum semantic similarity score is selected from the candidates as the final similarity between two words. Formally, given two potentially polysemous words w_1 and w_2, the semantic similarity between

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 2. *Statistics of the semantic taxonomy of noun synsets in WN*

Statistic	Value	Statistic	Value
Nodes	82,192	Edges	84,505
Number of inner nodes	17,161	Number of leaves	65,031
Maximum of depth	20	Maximum of density	414

Table 3. *Evaluation datasets of words*

Dataset	Collected by	Pair	Score	Subject
RG-65	Rubenstein and Goodenough (1965)	65	0–4	51
MC-30	Miller and Charles (1991)	30	0–4	31
WS-203	Agirre *et al.* (2009)	203	0–10	13–16
SimLex-666[1]	Hill, Reichart and Korhonen (2015)	666	0–10	50

them is computed as follows:

$$\mathrm{Sim}(t_1, t_2) = \max_{i,j} \ \mathrm{sim}\ \left(c_{1i}, c_{2j}\right) \tag{12}$$

where c_{1i} and c_{2j} are the concepts extracted from WN for the words t_1 and t_2, respectively.

4 Evaluation

This section describes in detail the evaluation of the proposed semantic similarity measure using four weighting mechanisms. This section also evaluates feature-based measures that depend on direct semantic relations and hierarchical feature-based measures that rely on taxonomical features when measuring the semantic similarity between words.

4.1 Experimental setup

In the experiments, the noun database in WN[2] 3.1 is used to build the semantic taxonomy to model the semantic measures. The database contains 82,192 noun synsets that have 117,947 unique terms (single and multiple words). Table 2 presents the descriptions of the semantic taxonomy using the noun synsets in WN.

In measuring the semantic relatedness between word pairs, three English gold-standard datasets (Mc-30, ReG-65, and WS-203) that have been widely used in several studies as well as the most recent dataset (SimLex-666) are employed to evaluate the semantic measures in the experiments in this work. Table 3 shows the description of these evaluation datasets.

The first dataset is RG-65 (Rubenstein and Goodenough 1965) that contains sixty-five word pairs ranging from 'semantically unrelated' to 'highly synonymous.'

[1] http://www.cl.cam.ac.uk/~fh295/simlex.html
[2] https://wordnet.princeton.edu/wordnet/download/current-version/

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Semantic similarity judgments have been obtained from fifty-one human subjects in a scale from 0 to 4. Miller and Charles (1991) selected a subset of thirty word pairs (MC-30) from the RG-65 dataset and obtained human judgments from thiry-one test subjects. The third dataset is WS-203 introduced by Agirre *et al.* (2009). WS-203 is made up of 203 word pairs and is a subset of the Fin-353 dataset. Fin-353 dataset (Finkelstein *et al.* 2001) contains 353 word pairs that have been rated for semantic relatedness by an average of 13–16 human annotators. Hill *et al.* (2015) introduced the largest and most recent word similarity benchmark (SimLex-999) in literature; it consists of 999 word pairs and comprises 666 noun–noun pairs, 222 verb–verb pairs, and 111 adjective–adjective pairs. Therefore, the final dataset is SimLex-666, which is a subset of SimLex-999 containing only noun–noun pairs.

4.2 Evaluation metrics

The *in-vitro* (direct) evaluation method (Zhang *et al.* 2013) is utilized in this study to evaluate the performance of the proposed method in measuring semantic similarity. In this method, the semantic similarity value of concepts is compared directly against a gold-standard dataset in terms of correlations. Two correlation coefficients are used extensively in the *in vitro* evaluation method; these two are Pearson's product–moment correlation r and Spearman's rank correlation ρ. These correlation coefficients are defined between the list of human judgments (X) and the list of values for the semantic relatedness measure (Y) and expressed as the following formulae:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{13}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{14}$$

where d_i is the difference between the ranks of x_i and y_i.

In Gabrilovich and Markovitch (2009), Hassan (2011), Sánchez and Batet (2012), Taieb *et al.* (2013), and Zhang *et al.* (2013), evaluation results were reported using Pearson, Spearman, or both. The Pearson coefficient imposes constraints on a correlation by imposing a linear relation between the two sets of normally distributed values being compared. The Spearman (nonparametric) coefficient relaxes this condition and only constrains the consistency of the relative ranking across the two inputs while imposing no distributional requirements. Hassan (2011) demonstrated that the harmonic mean of Pearson and Spearman is a good metric that can cover the linear relationship with human judgment and can satisfy the relative order imposed by these judgments. Following this researcher, the evaluation results are reported using the individual metric with the harmonic mean that is defined as follows:

$$\mu = \frac{2r\rho}{r + \rho} \tag{15}$$

where r and ρ are Pearson's product–moment correlation and Spearman's rank correlation, respectively.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 4. *Evaluation results of the proposed measure*

Datasets	Correlations	Edge	Depth	Density	Descendants
MC-30	r	0.802	0.813	0.865	0.852
	ρ	0.750	0.762	0.791	0.784
	μ	0.775	0.787	**0.826**	0.817
RG-65	r	0.829	0.860	0.859	0.862
	ρ	0.793	0.795	0.775	0.808
	μ	0.811	0.826	0.815	**0.834**
WS-203	r	0.642	0.678	0.702	0.724
	ρ	0.640	0.662	0.670	0.688
	μ	0.641	0.670	0.686	**0.706**
SimLex-666	r	0.595	0.598	0.592	0.600
	ρ	0.600	0.613	0.617	0.623
	μ	0.597	0.605	0.604	**0.611**

The results rendered in bold indicate the best results in terms of μ correlation in each dataset.

4.3 Evaluation results

The first experiment aims to evaluate the proposed measure for measuring semantic similarity between words. Given that the proposed measure depends on the weighting techniques of features in semantic representation, it is evaluated in this study using these techniques in four datasets. Table 4 summarizes the evaluation results for the proposed measure based on four weighting techniques using four topological parameters, namely, edge, depth, density, and descendants.

The second experiment is conducted to compare the proposed measure with hierarchical feature-based measures to determine the effect of the weights of features on semantic representation. In hierarchical feature-based measures, the concept is represented as a set containing its super-concepts in the semantic taxonomy. Based on set-based theory, the similarity between two sets s_1 and s_2 can be measured using the following functions.

$$\mathrm{Sim}_1(s_1, s_2) = |s_1 \cap s_2| \tag{16}$$

$$\mathrm{Sim}_2(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1| + |s_2|} \tag{17}$$

$$\mathrm{Sim}_3(s_1, s_2) = \frac{|s_1 \cap s_2|}{\max\left(|s_1|, |s_2|\right)} \tag{18}$$

$$\mathrm{Sim}_4(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \tag{19}$$

$$\mathrm{Sim}_{\mathrm{Batet}}(c_1, c_2) = -\log_2 \frac{|s_1 \cup s_2| - |s_1 \cap s_2|}{|s_1 \cup s_2|} \tag{20}$$

In Table 5, the evaluation results of hierarchical feature-based measures are presented for measuring the semantic similarity of words in the four datasets. $\mathrm{Sim}_{\mathrm{Batet}}(c_1, c_2)$ refers to the semantic measure introduced by Batet *et al.* (2011)

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 5. *Evaluation results of hierarchical feature-based measures*

Datasets	Correlations	Sim_1	Sim_2	Sim_3	Sim_4	Sim_{Batet}	Weighting based
MC-30	r	0.712	0.750	0.748	0.791	0.755	**0.852**
	ρ	0.695	0.758	0.751	0.758	0.758	**0.784**
	μ	0.703	0.754	0.749	0.774	0.756	**0.817**
RG-65	r	0.690	0.786	0.789	0.834	0.779	**0.862**
	ρ	0.671	0.784	0.784	0.784	0.784	**0.808**
	μ	0.680	0.785	0.786	0.808	0.781	**0.834**
WS-203	r	0.582	0.614	0.626	0.659	0.624	**0.724**
	ρ	0.579	0.652	0.645	0.652	0.652	**0.688**
	μ	0.580	0.632	0.635	0.655	0.638	**0.706**
SimLex-666	r	0.378	0.555	0.535	0.583	0.551	**0.600**
	ρ	0.358	0.578	0.534	0.578	0.578	**0.623**
	μ	0.368	0.566	0.534	0.580	0.564	**0.611**

In each row, the best correlation value is shown in bold.

to determine similarity based on the normalized ratio between the amount of uncommon features and the sum of common and uncommon features. The last column presents the results of the proposed measure using the descendants weighting technique.

The third experiment aims to compare feature-based measures with hierarchical feature-based measures and weighting-based measures. The synset can be represented as a set of features using direct relations in WN for measuring semantic similarity. In this experiment, the taxonomical relations (i.e., hypernyms and hyponyms) and non-taxonomical relations (e.g., substance meronym and member of domain usage) are used as a set of representative features to measure the semantic similarity between concepts. Similar to the procedure for hierarchical feature-based measures, (16)–(19) are used to measure the semantic similarity between two sets (representative features). The Sim_{Batet} measure is excluded because it was proposed to estimate the semantic similarity between hierarchical features. Table 6 presents a summary of the overall evaluation results for different representative features (direct hypernyms, direct hyponyms, combination of hypernyms with hyponyms, non-taxonomical relations, and combination of all direct relations in WN) using four similarity measures (referred to as Sim_1–Sim_4). In addition, this table presents the results of algorithms that were introduced in Saif *et al.* (2017) and Saif *et al.* (2014). The last two rows of the table show the best results obtained in hierarchical feature-based and weighting-based measures.

4.4 Statistical significance of the results

In statistical hypothesis testing, statistical significance plays an important role in determining whether the null hypothesis should be rejected or retained. This significance is measured using the p value, which is a number between 0 and 1, describing the probability of rejecting the null hypothesis when it is true. In most

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 6. *Evaluation results of feature-based measures*

Features	Measures	MC-30			RG-65			WS-203			SimLex-666		
		r	ρ	μ	r	ρ	μ	r	ρ	μ	r	ρ	μ
Hyper	Sim_1	0.775	0.799	0.787**	0.761	0.802	0.781	0.592	0.674	0.630	0.436	0.501	0.466
	Sim_2	0.775	0.799	0.787**	0.777	0.803	0.790	0.587	0.673	0.627	0.437	0.501	0.467
	Sim_3	0.775	0.799	0.787**	0.777	0.803	0.790	0.585	0.673	0.626	0.437	0.501	0.467
	Sim_4	0.707	0.799	0.750	0.721	0.803	0.760	0.545	0.673	0.602	0.433	0.501	0.465
Hypo	Sim_1	0.399	0.807	0.534	0.391	0.799	0.525	0.244	0.623	0.351	0.254	0.572	0.352
	Sim_2	0.609	0.799	0.691	0.630	0.800	0.705	0.403	0.624	0.490	0.370	0.574	0.450
	Sim_3	0.572	0.798	0.667	0.600	0.800	0.686	0.382	0.624	0.474	0.352	0.573	0.436
	Sim_4	0.567	0.799	0.663	0.596	0.800	0.683	0.378	0.624	0.471	0.347	0.574	0.433
Hyper $\cup$ Hypo	Sim_1	0.450	0.807	0.578	0.459	0.799	0.583	0.333	0.685	0.448	0.310	0.564	0.400
	Sim_2	0.601	0.794	0.684	0.633	0.799	0.707	0.477	0.679	0.560	0.337	0.522	0.410
	Sim_3	0.618	0.798	0.697	0.638	0.800	0.710	0.477	0.679	0.561	0.370	0.523	0.434
	Sim_4	0.609	0.799	0.691	0.630	0.800	0.705	0.459	0.680	0.548	0.376	0.525	0.438
Non-taxonomical	Sim_1	0.316	0.622	0.419	0.287	0.658	0.400	0.211	0.577	0.308	0.191	0.382	0.255
	Sim_2	0.524	0.620	0.568	0.570	0.661	0.612	0.397	0.579	0.471	0.253	0.386	0.305
	Sim_3	0.520	0.619	0.565	0.568	0.662	0.611	0.393	0.579	0.468	0.263	0.386	0.313
	Sim_4	0.509	0.620	0.559	0.559	0.661	0.606	0.384	0.579	0.462	0.266	0.386	0.315
All	Sim_1	0.367	0.834	0.509	0.347	0.814	0.486	0.259	0.700	0.378	0.277	0.490	0.354
	Sim_2	0.592	0.820	0.687	0.548	0.807	0.653	0.476	0.695	0.565	0.219	0.460	0.297
	Sim_3	0.599	0.829	0.696	0.628	0.812	0.708	0.472	0.696	0.563	0.350	0.475	0.403
	Sim_4	0.599	0.829	0.695	0.628	0.812	0.708	0.472	0.696	0.563	0.345	0.475	0.400
Saif *et al.* (2014)		0.732	0.695	0.713	0.763	0.708	0.735	0.640	0.618	0.629	0.491	0.493	0.492
Saif *et al.* (2017)		0.722	0.737	0.729	0.780	0.788	0.784	0.599	0.639	0.618	0.548	0.574	0.561
Hierarchical feature-based measures		0.791	0.758	0.774	0.834	0.784	0.808**	0.659	0.652	0.655**	0.583	0.578	0.580**
Weighting-based		0.852	0.784	0.817*	0.862	0.808	0.834*	0.724	0.688	0.706*	0.600	0.623	0.611*

*First rank. **Second rank.

Table 7. *Statistical significance of the difference between the proposed measure and human judgments*

Parameters	MC-30	RG-65	WS-203	SimLex-666
Edge	9.77E-08*	1.41E-17*	7.69E-24*	3.75E-65*
Depth	4.68E-08*	4.38E-20*	2.37E-27*	9.09E-66*
Density	7.07E-10*	5.09E-20*	5.05E-30*	1.14E-61*
Descendants	2.25E-09*	3.05E-20*	1.08E-32*	2.64E-66*

The results marked with an asterisk indicate statistically significant improvements in the weighting-based results at the 0.05 level of significance.

empirical analyses, results yielding a p value of 0.05 are considered on the borderline of statistical significance. Results are considered statistically significant when the p value is less than 0.01. The p values less than 0.005 indicate that the results are considered highly statistically significant. In measuring semantic similarity, two cases of statistical significance are used to compare two groups. The first case involves a comparison between human judgments and machine similarity values. The second case involves a comparison between the proposed measure and the baseline measure. In the first case, the null hypothesis is that the scores of similarity in a gold-standard dataset and the computed similarity values are unrelated. From the correlation coefficient and t test, the p value in this case, as shown in Table 7, is computed using a software introduced by Wessa (2016).

In the second case, Steiger's technique with z test[3] (Steiger 1980) that was introduced for dependent correlations is used to measure the statistical significance of the difference between the performance of the proposed method and that of the baseline methods. Given that the proposed measure and each baseline method compute the correlation coefficients against the same test set, Steiger's technique is more suitable than Fisher's z transformation that is used for independent correlation coefficients, as presented in Gabrilovich and Markovitch (2009). We let X be the human scores in the test set, Y be the similarity results obtained by the proposed method, and Z be the results obtained by the baseline method. Steiger's technique is employed to examine whether the correlation score XY (representing the accuracy of the proposed method) is significantly different from XZ (representing the accuracy of the baseline method) in consideration of the correlation score YZ. The statistical significance (two-tailed p value) of the difference between the Pearson's correlation of the proposed measure and that of the baseline measures (Sim_1 and Sim_{Batet}) in the evaluation test sets is presented in Tables 8 and 9.

4.5 Discussion

The evaluation results of the three experiments are analyzed for three purposes. The first one is to show the effect of exploiting topological parameters in a semantic

[3] The online calculator of this technique is available at http://quantpsy.org/corrtest/corrtest2.htm.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 8. *Statistical significance of the difference between the proposed measure and the baseline method* (Sim_1)

Parameters	MC-30	RG-65	WS-203	SimLex-666
Edge	8.76E-02	2.50E-05*	1.82E-02*	3.11E-36*
Depth	5.03E-02	7.00E-06*	9.90E-05*	1.14E-34*
Density	6.75E-04*	8.00E-06*	2.49E-09*	1.51E-32*
Descendants	2.85E-03*	2.00E-06*	7.51E-13*	2.87E-48*

The results marked with an asterisk indicate statistically significant improvements in the weighting-based results at the 0.05 level of significance.

Table 9. *Statistical significance of the difference between the proposed measure and* Sim_{Batet}

Parameters	MC-30	RG-65	WS-203	SimLex-666
Edge	4.47E-01	1.16E-01	2.94E-01	4.08E-03*
Depth	3.25E-01	4.12E-03*	2.12E-02*	7.71E-03*
Density	3.15E-02*	1.37E-02*	1.41E-04*	3.78E-03*
Descendants	8.54E-02	1.73E-03*	7.66E-08*	1.34E-03*

The results marked with an asterisk indicate statistically significant improvements in the weighting-based results at the 0.05 level of significance.

taxonomy in hierarchical feature-based measures. The second aim is to compare the weighting schemes in terms of measuring semantic similarity. The final aim is to confirm the superior performance of hierarchical feature-based measures compared with other feature-based measures.

Based on the evaluation results shown in Tables 4 and 5, Figure 5 shows a comparison between the proposed measure with four weighting techniques and five hierarchical feature-based measures in four datasets. The proposed measure using three weighting techniques (i.e., depth, density, and descendants) outperforms all hierarchical feature-based measures in all the datasets. This finding is expected because the topological-based measure was proposed to address the issue of the constant weighting assumption in hierarchical feature-based measures. The proposed measure overcomes the main limitation of hierarchical feature-based measures that do not consider the specificity of features in semantic representation. This finding strengthens our motivation to use topological parameters in representing the semantics of concepts instead of using only unweighted features, as in hierarchical feature-based measures.

The comparison of the weighting techniques of the proposed measure reveals that the descendant-based weighting technique achieved the highest μ correlation in three datasets (RG-65, WS-203, and SimLex-666). This finding can be attributed to the fact that this technique depends on the generic–specific relations that have been formalized in terms of the logic of classes. In MC-30, the density-based technique

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

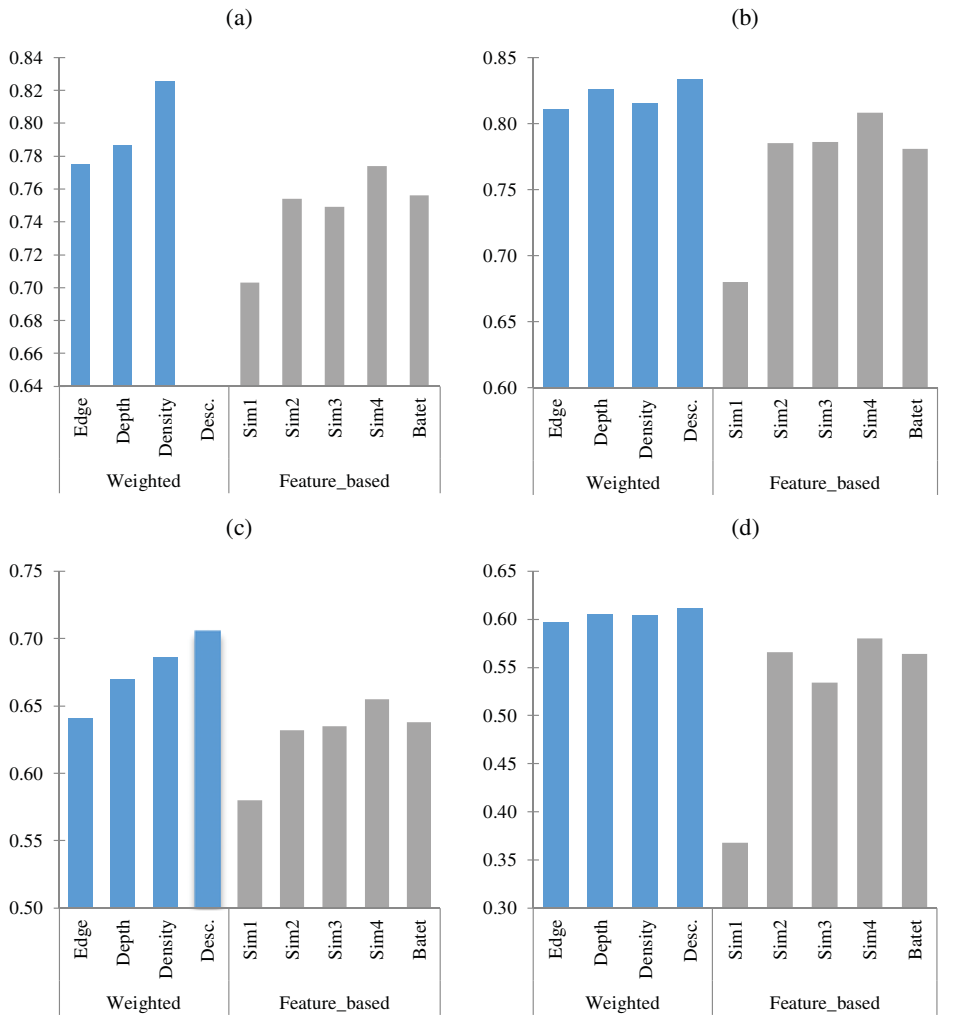


Fig. 5. (Colour online) Comparison between the proposed measure and hierarchical feature-based measures in (a) MC-30, (b) RG-65, (c) WS-203, and (d) SimLex-666 datasets.

achieves the highest μ correlation (0.826) among the other techniques (0.775, 0.787, and 0.817).

In Table 4, the semantic measures that depend on hypernyms achieve the highest μ correlations among those that depend on hyponyms, the combination of hypernyms with hyponyms, nontaxonomical relations, and all relations. For taxonomical relations, the hypernyms tend to be more useful when measuring semantic similarity than the hyponyms. This finding can be attributed to the fact that the number of hyponyms in many concepts is higher than the number of hypernyms. This finding indicates that the semantic similarity between two words is affected by the number of hyponyms. For example, the concept 'voyage' is the hyponym of the concept 'journey,' but the semantic similarity between them using

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

hyponyms is 0.052 because the union of the hypernyms of these two concepts contains nineteen concepts. By contrast, the union of the hypernyms of these two concepts is 2, leading to the 0.5 similarity between them. For this reason, semantic representation using the combination of hypernyms with hyponyms obtains lower correlation than using hypernyms only.

For nontaxonomical relations, the similarity measures achieve lower correlation than the hypernym-based measures because non-taxonomical relations are considered useful when measuring semantic relatedness. Given that semantic representation using all relations may contain more semantic evidence on the associations between concepts, the experimental results show that the performance of the semantic measures using this representation is poorer than that of the semantic measures using hypernyms only. However, the semantic measures (Saif *et al.* 2014; Saif *et al.* 2017) that depend on the selected features from all relations achieve competitive results as the hypernym-based measures and outperform the hypernym-based measures in the largest dataset (SimLex-666).

4.6 Comparison with IC-based measures

IC-based measures have been introduced for modeling the semantic similarity between two lexical concepts based on the information they share in common. These measures are classified according to the type of sources which are used for computing the information content of the concepts into two methods: corpus-based and ontology-based (Intrinsic) methods. The corpus-based method requires a tagged corpus for computing the IC of a concept, while the intrinsic method exploits the taxonomic parameters to determine the IC of a concept. In this section, to ensure a fair comparison of the results, the proposed method is compared with only the intrinsic IC-based method which shares some aspects (using the taxonomical parameters) with the proposed method.

The taxonomical parameters are used in path-based measures to address the specificity issue and in IC measures to determine the IC of concepts instead of using the tagged corpus. The results obtained by the proposed measure using four weighting techniques are compared with those obtained by other intrinsic IC-based similarity measures to investigate the difference between using the parameters in IC measures and the proposed measure. Lastra-Díaz and García-Serrano (2015a) presented a comprehensive evaluation and comparison of several IC measures. In the datasets that contain a large number of word pairs (WS-203 and SimLex-666), Table 10 shows the best result obtained for IC-based similarity measures using different settings that produce the best value for each measure as well as the results obtained by the proposed measure.

The results achieved in terms of μ correlation by the proposed measure when using the descendant-based weighting technique are the highest (0.706 and 0.611 in WS-203 and SimLex-666, respectively) compared with the best results obtained from the family of IC-based measures. Although these IC-based measures use the same taxonomical parameters to assess similarity, the proposed measure better captures semantic evidence on the semantic association between words, thus enabling a

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Table 10. *Comparison of IC-based measures and the proposed measure*

Proposed by	Evaluated in	WS-203			SimLex-666		
		r	ρ	μ	r	ρ	μ
Yuan *et al.* (2013)	Lastra-Díaz and García-Serrano (2015a)	0.706	0.666	0.685	0.611	0.603	0.607
Seco *et al.* (2004)		0.697	0.664	0.680	0.605	0.590	0.597
Sánchez *et al.* (2011)		0.695	0.658	0.676	0.594	0.591	0.592
Meng *et al.* (2012)		0.704	0.658	0.680	0.601	0.596	0.598
Harispe *et al.* (2014)		0.696	0.654	0.674	0.606	0.592	0.599
Sánchez *et al.* (2012)		0.697	0.659	0.678	0.600	0.585	0.592
Lastra-Díaz and García-Serrano (2015a)		0.690	0.652	0.671	0.596	0.583	0.590
Zhou *et al.* (2008)		0.685	0.659	0.672	0.599	0.595	0.596
Taieb, Aouicha and Hamadou (2014)		0.649	0.618	0.633	0.492	0.483	0.488
Edge	This work	0.642	0.64	0.641	0.595	0.600	0.597
Depth		0.678	0.662	0.670	0.598	0.613	0.605
Density		0.702	0.670	0.686	0.592	0.617	0.604
Descendants		0.724	0.688	**0.706**	0.600	0.623	**0.611**

The results rendered in bold indicate the best results in terms of μ correlation in each dataset.

more accurate differentiation of words and a more precise similarity measurement. This finding can be attributed to the fact that the proposed measure considers the weighting of shared and unshared features between two concepts instead of the weighting of the two concepts with their least common subsumer in IC-based measures.

Based on comparative results, it is important to note that the improvement of the proposed method against IC-based measures is less noticeable, especially for the larger dataset. Previous experimental results (Lastra-Díaz and García-Serrano 2015a; Aouicha *et al.* 2016) have shown a similar observation when comparing among taxonomical IC-based measures. For the WS-203 dataset, the improvement ratio (using μ correlation) of the best IC-based measure (Yuan *et al.* 2013) is 0.005 against the earlier taxonomical IC-based measure (Seco *et al.* 2004). This is largely due to that the taxonomical measures depend mainly on structural features which have limitations in terms of completeness and balancing on granularity. For instance, based on statistics of WordNet, around 79 per cent of concepts have the same descendant (equal to zero) which means that these concepts have the same degree of specificity.

In terms of the performance of WN-based semantic measures, it could be observed that the accuracy of these measures on WS-203 and SimLex-666 datasets is limited, since the best accuracy values (in terms of μ correlation) are 0.611 and 0.706 on these datasets, respectively. This means that the semantic similarity scores of human judgments and WN-based measures did not correlate well on the WS-203 and SimLex-666 datasets. This can be interpreted using two reasons which are related

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

to the knowledge source (WordNet) and the datasets used in the evaluation. The first reason is the limitations of the WN in terms of the coverage and completeness of the semantic relations. Although the WN has been widely used for measuring semantic similarity in the literature, it still requires updating to enrich its semantic relations. The second reason is the difference in the granularity of semantics between the evaluation datasets (the coarse-grained annotation) and the WN (the fine-grained annotation). In the coarse-grained annotation, the annotators pay more attention to the general concepts for quantifying the semantic similarity between two words. On the SimLex-666 dataset, there is a number of word pairs which are synonyms (the semantic similarity between each word pair equals one) in WN and have been annotated by different human judgment scores. For example, the word pairs 'door-doorway,' 'molecule-atom,' 'mouth-lip,' 'couple-pair,' and 'teacher-instructor' have been annotated to have the semantic similarity scores with a [0–1] scale: 0.540, 0.645, 0.710, 0.833, and 0.925, respectively. However, all WN-based semantic measures return the same score (1, the highest score) for all synonyms.

5 Conclusion

This study proposed a new weighting-based semantic similarity measure that overcomes the main limitation of hierarchical feature-based measures. From the semantic taxonomy of the lexical source, the semantics of a concept were represented by the features from its super-concepts. Then, the topological parameters were exploited to introduce four weighting techniques (edge, depth, density, and descendants). Given that the proposed method requires only the taxonomy, the proposed method can be applied to any semantic taxonomy or ontology to measure the similarity between two nodes in the taxonomy. The performance of the proposed measure in measuring the semantic relatedness of words was evaluated by using the semantic taxonomy of WN. The experimental results indicated that the proposed measure could overcome the limitations of hierarchical feature-based measures, given that the correlation of the proposed measure was higher than that of feature-based measures in different evaluation datasets. Comparison with other semantic similarity paradigms showed that the proposed measure outperformed the IC-based measures because the proposed measure provided higher correlations with human judgments than other IC-based measures when applied to the largest datasets of semantic similarity measurement. In fact, better correlations ($\mu = 0.706$) were achieved in the WS-203 dataset, whereas the best correlations ($\mu = 0.685$) were achieved in all IC-based measures.

The proposed measure can be utilized to exploit other taxonomical parameters or to integrate a number of parameters to achieve a more robust weighting technique. The proposed taxonomy-based similarity measure can be applied to several ontologies for specific domains, such as MeSH or SNOMED for the biomedicine domain or gene ontology for the biomedical domain that can be formed as taxonomies, and for ontologies of other knowledge sources, such as Wikipedia.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, USA. Association for Computational Linguistics, pp. 19–27.

Al-Mubaid, H., and Nguyen, H. A. 2006. A cluster-based approach for semantic similarity in the biomedical domain. In *Proceedings of the 28th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, 2006*. New York, USA, pp. 2713–7.

Aouicha, M. B., Taieb, M. A. H., and Ezzeddine, M. 2016. Derivation of "is a" taxonomy from Wikipedia category graph. *Engineering Applications of Artificial Intelligence* **50**: 265–86.

Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the *18th* International Joint Conference on Artificial Intelligence, Acapulco, Mexico, pp. 805–10.

Batet, M., Sánchez, D., and Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics* **44**(1): 118–25.

Cross, V., Yu, X., and Hu, X. 2013. Unifying ontological similarity measures: A theoretical and empirical investigation. *International Journal of Approximate Reasoning* **54**(7): 861–75.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2001. Placing search in context: The concept revisited. In Proceedings of the 10th International Conference on World Wide Web, Hong Kong, ACM, pp. 406–14.

Firth, J. R. 1957. *A Synopsis of Linguistic Theory, 1930–1955*. In Studies in Linguistic Analysis. Oxford: Blackwell.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, Morgan Kaufmann, pp. 1606–11.

Gabrilovich, E., and Markovitch, S. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* **34**(2): 443–98.

Gentleman, R. 2005. Visualizing and distances using GO. `http://www.bioconductor.org/docs/vignettes.html`.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. 2007. Topics in semantic representation. *Psychological Review* **114**(2): 211–44.

Gurevych, I. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Natural Language Processing–IJCNLP* 2005, pp. 767–78. Berlin: Springer.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J. 2014. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics* **48**: 38–53.

Hassan, S. 2011. *Measuring Semantic Relatedness Using Salient Encyclopedic Concept*. PhD thesis, University of North Texas, Denton, TX, USA.

Hassan, S., and Mihalcea, R. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, Association for Computational Linguistics, pp. 1192–201.

Hassan, S., and Mihalcea, R. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of AAAI 2011 (25th AAAI Conference on Artificial Intelligence)*, San Francisco, Association for the Advancement of Artificial Intelligence, pp. 884–9.

Hill, F., Reichart, R., and Korhonen, A. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* **41**(4): 665–95.

Jiang, J. J., and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING 97)*, Taiwan, pp. 19–33.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Jiang, Y., Bai, W., Zhang, X., and Hu, J. 2016. Wikipedia-based information content and semantic similarity computation. *Information Processing & Management* **53**(1), 248–65.

Jiang, Y., Zhang, X., Tang, Y., and Nie, R. 2015. Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing & Management* **51**(3): 215–34.

Lastra-Díaz, J. J., and García-Serrano, A. 2015a. A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems* **89**: 509–26.

Lastra-Díaz, J. J., and García-Serrano, A. 2015b. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence* **46**: 140–53.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual International Conference on Systems Documentation*, Toronto, Canada, ACM, pp. 24–6.

Meng, L., Gu, J., and Zhou, Z. 2012. A New model of information content based on concept's topology for measuring semantic similarity in WordNet. *International Journal of Grid & Distributed Computing* **5**(3): 81–94.

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

Pirró, G. 2009. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering* **68**(11): 1289–308.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, ACM, pp. 337–46.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* **8**(10): 627–33.

Saif, A., Ab Aziz, M. J., and Omar, N. 2014. Evaluating knowledge-based semantic measures on Arabic. *International Journal on Communications Antenna and Propagation* **4**(5): 180–94.

Saif, A., Ab Aziz, M. J., and Omar, N. 2016. Reducing explicit semantic representation vectors using Latent Dirichlet Allocation. *Knowledge-Based Systems* **100**: 145–59.

Saif, A., Ab Aziz, M. J., and Omar, N. 2017. Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features. *Natural Language Engineering* **23**(1): 53–91.

Sánchez, D., and Batet, M. 2012. A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and Information Systems (IJSWIS)* **8**(2): 34–50.

Sánchez, D., Batet, M., and Isern, D. 2011. Ontology-based information content computation. *Knowledge-Based Systems* **24**(2): 297–303.

Sánchez, D., Batet, M., Isern, D., and Valls, A. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications* **39**(9): 7718–28.

Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI 2004, Including Prestigious Applicants of Intelligent Systems*, Valencia, Spain, IOS Press, pp. 1089–90.

Steiger, J. H. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* **87**(2): 245–51.

Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the 2nd International Conference on Information and Knowledge Management*, Washington, D.C., USA: ACM, pp. 67–74.

Taieb, H., Ben Aouicha, M., Tmar, M., and Hamadou, A. B. 2011. New information content metric and nominalization relation for a new WordNet-based method to measure the semantic relatedness. In *IEEE 10th International Conference on, Cybernetic Intelligent Systems (CIS)*, 2011, London, UK: IEEE, pp. 51–8.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press

Taieb, M. A., Ben Aouicha, M., and Ben Hamadou, A. 2013. Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems* **50**: 260–78.

Taieb, M. A. H., Aouicha, M. B., and Hamadou, A. B. 2014. Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence* **36**: 238–61.

Wang, T., and Hirst, G. 2011. Refining the notions of depth and density in wordnet-based semantic similarity measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, UK, pp. 1003–11.

Wessa, P. 2016. Free Statistics Software, Office for Research Development and Education. `http://www.wessa.net/corr.wasp`, Accessed November 9, 2016.

Wu, X., Zhu, L., Guo, J., Zhang, D.-Y., and Lin, K. 2006. Prediction of yeast protein–protein interaction network: Insights from the gene ontology and annotations. *Nucleic Acids Research* **34**(7): 2137–50.

Yuan, Q., Yu, Z., and Wang, K. 2013. A new model of information content for measuring the semantic similarity between concepts. In *International Conference on Cloud Computing and Big Data (CloudCom-Asia), 2013*, Fuzhou, China: IEEE, pp. 141–6.

Zhang, Z., Gentile, A. L., and Ciravegna, F. 2013. Recent advances in methods of lexical semantic relatedness–a survey. *Natural Language Engineering* **19**(4): 411–79.

Zhou, Z., Wang, Y., and Gu, J. 2008. A new model of information content for semantic similarity in WordNet. In *Proceedings of the 2nd International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*, Hainan Island, China, IEEE, pp. 85–9.

https://doi.org/10.1017/S1351324918000190 Published online by Cambridge University Press