Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-021-00248-0

REGULAR PAPER

Knowledge-based sentence semantic similarity: algebraical properties

Mourad Oussalah[1] · **Muhidin Mohamed**[2]

Received: 19 October 2019 / Accepted: 1 June 2021

Abstract
Determining the extent to which two text snippets are semantically equivalent is a well-researched topic in the areas of natural language processing, information retrieval and text summarization. The sentence-to-sentence similarity scoring is extensively used in both generic and query-based summarization of documents as a significance or a similarity indicator. Nevertheless, most of these applications utilize the concept of semantic similarity measure only as a tool, without paying importance to the inherent properties of such tools that ultimately restrict the scope and technical soundness of the underlined applications. This paper aims to contribute to fill in this gap. It investigates three popular WordNet hierarchical semantic similarity measures, namely path-length, Wu and Palmer and Leacock and Chodorow, from both algebraical and intuitive properties, highlighting their inherent limitations and theoretical constraints. We have especially examined properties related to range and scope of the semantic similarity score, incremental monotonicity evolution, monotonicity with respect to hyponymy/hypernymy relationship as well as a set of interactive properties. Extension from word semantic similarity to sentence similarity has also been investigated using a pairwise canonical extension. Properties of the underlined sentence-to-sentence similarity are examined and scrutinized. Next, to overcome inherent limitations of WordNet semantic similarity in terms of accounting for various Part-of-Speech word categories, a WordNet "All word-To-Noun conversion" that makes use of Categorial Variation Database (CatVar) is put forward and evaluated using a publicly available dataset with a comparison with some state-of-the-art methods. The finding demonstrates the feasibility of the proposal and opens up new opportunities in information retrieval and natural language processing tasks.

Keywords Sentence semantic similarity · Part-of-speech conversion · WordNet · CatVar

1 Introduction

Measures of semantic similarities have been primarily developed for quantifying the extent of resemblance between two words or two concepts using pre-existing resources that encode word-to-word or concept-to-concept relationships as in WordNet lexical database [1, 2].

Accurate comparison between text snippets for the similarity determination is a fundamental prerequisite in the areas of natural language processing, information retrieval, text summarization, document clustering, question answering, automatic essay scoring and others [3, 4]. For instance, the quantification of the similarity between candidate sentences can allow us to promote a good summary coverage and prevent redundancy in automatic text summarization [5]. In this respect, the similarity values of sentence pairs are sometimes used as part of the statistical features of the text summarization system. Likewise, question answering applications require similarity identification between a question–answer or question–question pair [6]. Similarly, query-based summarization crucially relies on similarity scores for summary extraction [7]. On the other hand, semantic similarity plays a crucial role in information retrieval where the matching documents are ranked according to their similarity with the supplied query [3, 8, 9]. Plagiarism detection is a recent area of research which is solely based on text similarity detection [10–12]. Among other text similarity applications, one shall mention machine translation [13], text classification [3, 4, 14], database where similarity is used for schema matching [15], and bioinformatics [16].

✉ Mourad Oussalah
mourad.oussalah@oulu.fi

[1] Faculty of Information Technology and Electrical Engineering, CMVS, University of Oulu, 90014 Oulu, Finland

[2] Operations and Information Management Department, Aston University, Birmingham, UK

Published online: 21 August 2021

Strictly speaking, computing sentence similarity is not trivial due to the variety of linguistic constructs and inherent ambiguity in textual expressions, which renders the task of determining semantically equivalent sentences very challenging even for human beings, especially when the contextual information is not well-known [4].

Conventionally, sentence-to-sentence similarity computation can be categorized into two streams of approaches. The first one advocates a lexical or a string-based approach where sentence similarity quantifies the amount of overlapping of the characters constituting the two input sentences. A such approach is primarily intended to capture similarity among the input sentences in terms of the extent of the string sequence matching between the two inputs. This can provide insights to identify gaps that might be attributed to misspelling as in edit distance [17]. The second stream corresponds to the semantic similarity which attempts to capture the overlap between the meanings conveyed by the individual sentences [18]. This category includes several other approaches as well. Notably, one distinguishes (i) the corpus-based approach which uses statistical analysis about word co-occurrence over a large corpus as in the newly emerging distributional and embedding methods [19]; (ii) knowledge-based approach, which relies on a handcrafted semantic net of words as in WordNet lexical database [20]; (iii) feature-based methods where sentences are represented through a vector of predefined features [19]; hybrid-based approach that seeks to leverage features methods with corpus-based or knowledge-based approaches [18]. In parallel, one shall also mention the growing interest to deep learning architecture-based semantic similarity where attention weighs mechanism was proposed to accommodate fine-grain variations in various linguistic constructs as in Quan et al. [21]

In the last decade, knowledge-based linguistic measures with their various variants become the state-of-the-art methodology for computing the similarity scores among pairs of text snippets, which forms the basis of many commercial plagiarism detection tools, automatic summarization and information retrieval systems [4]. In essence, these methods utilize word-level semantic networks and lexical relations to arrive at sentence-level relevance. Electronic resources such as WordNet [1, 2] and ConceptNet [22] are acknowledged as the main lexical resources and knowledge bases for this purpose where the quantification of the semantic similarity between two sentences is evaluated as a global measure on pairwise comparison of word similarity of these sentences [23–26]. Nevertheless, a such construction of sentence-to-sentence semantic similarity from individual word semantic similarity is also prone to at least three inherent limitations and challenges.

- (A1) The order of words in individual sentences cannot be accounted for. Therefore, sentences like "students like fast exercises in class" and "fast students like class in" would yield the same semantic similarity. We refer to this challenge *A1*. Attempts to account for word order in the semantic similarity have been investigated by several researchers but with very limited success. For instance, Islam and Inkpen [27] suggested to include a word-order similarity as a weighted additive component to the semantic similarity, where the word-order similarity is computed by the normalized difference of common words among the tokens of two input sentences. This presupposes the existence of a set of tokens, which are common to both inputs. A very similar approach has also been adopted by Li et al. [25, 26]. Ozates et al. [28] proposed to use the dependency grammar concept. In essence their approach uses dependency tree bigram units and evaluates the similarity of the two input sentences as the amount of the bigram unit match. Nevertheless, it should be noted that handling bigram or n-gram unit instead of standard bag-of-word representation entails substantial increase of computational time, without necessarily achieving higher accuracy score as pointed out in some other studies [18, 20].
- (A2) The WordNet word-level semantic is restricted to noun and verb part-of-speech (PoS) categories only. This makes nouns and verbs as the two only classes usable to calculate the similarity scores between individual words because of their hierarchical taxonomic organization in WordNet. This trivially leaves the correlation among entities of distinct PoS as well as other types of non-verbal and non-naming entities uncounted for. For instance, WorldNet measures fail to establish the semantic relation of the word "investigate" to any of the words "investigation", "investigator", or "investigative" as they belong to different classes (part of speech) which are not taxonomically linked in WordNet hierarchy. Although the derivational morphology of words is contained in the WordNet lexical database, but in a distinct fashion and without explicit coherence, such limitations have already been pointed out in other studies, see, for instance [1, 2]. Therefore, accounting for various PoS entities in the quantification of the sentence-to-sentence semantic similarity remains an open challenge to the research community. We shall refer to this challenge *A2*.
- (A3) Several WordNet semantic similarity measures have been put forward by the researchers, some of which are solely based on the hierarchy of the WordNet taxonomy [29, 30], while others make use corpus-based information as well [31–33]. Therefore, the contribution of individual semantic similarity measure to sentence-to-sentence similarity is still to be investigated. Although several studies have reported some ad hoc and experimental-based

comparison as pointed out in [34–38], the almost absence of theoretical studies in this respect is quite striking. We shall refer to this challenge *A3*.

This paper aims to address challenges *A2* and *A3*. The cornerstone ingredients for addressing these challenges are twofold. To address challenge *A2*, we advocate a new proposal for PoS word conversation using WordNet taxonomy and derivational relations to promote the "All-to-noun" conversation, where all lexemes are turned into their noun counterparts. The all-to-noun conversion is motivated by several grounded arguments. First, noun entities are much more abundant in WordNet taxonomy than verbal or other PoS categories [2, 39]. Second, the graph hierarchy of noun entities is more elaborated than its verbal counterpart as testified by its larger depth. Third, it provides an appropriate framework to handle named entities. A testing experiment is put forward to quantify the usefulness of a such transformation. Next, to address challenge *A2*, we hypothesize that the knowledge of algebraical properties of semantic similarity measures would provide useful insights to choose appropriate semantic measure in a given context. Especially, we mainly focus on hierarchical-based semantic similarity measures, namely path-length measure [37], Wu and Palmer measure [29] and Leacock and Chodorow [30] measure because of their dominance in the text mining literature and applications [3, 40], where some theoretical results related to the expected performance of such measures are pointed out. Particularly, we examine the properties of these measures in terms of range, monotonicity, boundary and computational complexity. The extension from word-to-word to sentence-to-sentence semantic similarity is also investigated. Particularly, algebraical and intuitive properties of the sentence-to-sentence similarity measure when using the main WordNet hierarchical-based semantic similarity measures are laid bare, providing a tremendous help to a user when deciding to choose a particular measure according to his/her context. In terms of research questions, this paper tackles the following:

Q1 What is the benefit of the "All-to-noun" derivation transformation in the calculus of the sentence-to-sentence semantic similarity?

Q2 What are the algebraical properties of the main hierarchical WordNet semantic similarity measures?

Q3 How are these properties extended from word-level semantic similarity to sentence-level similarity?

The rest of the paper is organized as follows. Section 2 reviews WordNet semantic similarity measures where algebraical properties of the main taxonomy-based measures are examined, which tackle Research Question *Q2*. Section 3 investigates the sentence-to-sentence semantic similarities and their associated properties, contributing to Research Question *Q3*. In Sect. 4, we will explore the WordNet PoS conversions detailing the implication of such conversion on sentence-to-sentence semantic similarity measures through exemplification and testing. Especially, the performance of the suggested all-to-noun conversation in terms of sentence-to-sentence semantic similarity using a publicly available dataset is performed. The results were also compared to other conversion approaches, contributing to Research Question *Q1*.

2 WordNet and semantic similarity

2.1 WordNet taxonomy

In WordNet [1, 2], words are clustered into synsets, which are considered to be either synonym or at least carry the same semantic meaning. The words in each synset are grouped so that they are interchangeable in certain contexts. Typically, synsets are used to represent lexical concepts by bounding words and word senses together in a lexical matrix. A single word may appear in more than one synset, which agrees with the fact that a word may have multiple meanings or can appear in multiple parts of speech (verb, noun and adjective/adverb).

Synsets are linked with each other through various semantic relations. The most common relationships are the *Hyponym/Hypernym* and *Meronym/Holonym* relationships, which are defined for noun category. Hyponymy (resp. hypernym) describes the relation of being subordinate (resp. superior) or belonging to a lower (resp. higher) rank or a class. Meronymy–holonym corresponds to the relation that holds between a part and the whole. For verb category, *Troponym* and *Entailment* relations are examples of implemented hierarchical relationships in WordNet.

This design creates a sequence of levels going from a specific word (noun or verb) from a lower level to a broader category at the top level. In WordNet, the relation between lexicalized concepts is implemented by a pointer between the corresponding synsets. For example, if we use brackets to indicate synsets and "@ →" to represent a relation to the meaning of 'IS-A' or 'IS-A-KIND-OF' a possible hierarchy would be:

{robin, redbreast} @ → {bird} @ → {animal, animate_being} @ → {organism, life_form, living_thing}

This manner of representing hyponymy and hypernymy yields a lexical hierarchy in the form of a tree diagram. Such hierarchies are called inheritance systems because items are inheriting information from their superordinates. Thus, all properties of the superordinate are assumed to be properties of the subordinate object as well, while the nouns in WordNet are an example of a lexical inheritance system.

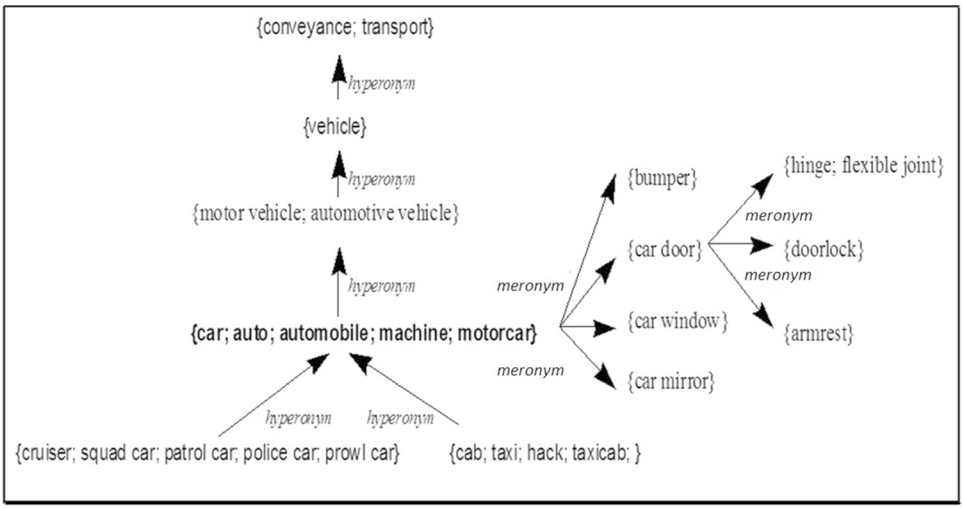

Fig. 1 Example of WordNet hierarchy

In theory, it is possible to combine all hypernyms into one hierarchy subordinate to an empty synset with no superordinates called a unique beginner. In WordNet, there are several noun hierarchies each starting with a different unique beginner. These multiple hierarchies belong to distinct semantic fields, each with a different vocabulary. Furthermore, since all hyponyms inherit the information of their hypernym, each unique beginner represents a primitive semantic component of the hyponym in question.

Unlike nouns, some verbs are not organized in hierarchies. These are connected by antonym relationships and similar relationships, like adjectives. Adjectives are organized into both head and satellite synsets, each organized around antonymous pairs of adjectives. These two adjectives are considered head synsets. Satellite synsets include words whose meaning is similar to that of a single head adjective. Nouns and adjectives that are derived from verbs and adverbs that are derived from adjectives have pointers indicating these relations.

An example of WordNet hierarchies is shown in Fig. 1 (taken from WordNet 1.5). In this example, the synset {car; auto; automobile; machine; motorcar} is related to:

- A broader concept (hypernym synset): {motor vehicle; automotive vehicle},
- More specific concepts or hyponym synsets: e.g. {cruiser; squad car; patrol car; police car; prowl car} and {cab; taxi; hack; taxicab},
- Parts of it is composed of: {bumper}; {car door}, {car mirror} and {car window} (meronymy relationship).

Now given a specific hierarchy of noun/verb concepts, one can estimate the semantic similarity between any two concepts by exploring the structure of the hierarchy between the two underlying concepts. This is detailed in the forthcoming sections, where a theoretical investigation of some of the key measures making use of only graph structure is carried out.

3 Taxonomy-based semantic similarity

We shall use the following notations:

- $len(c_i,c_j)$: the length of the shortest path from synset c_i to synset c_j in WordNet lexical database.
- $lcs(c_i,c_j)$: the lowest common subsumer of c_i and c_j
- $depth(c_i)$: shortest distance in terms of number of edges or nodes from global root to synset c_i, where the global root is such that $depth(\text{root})=1$.
- max_depth: the maximum depth of the taxonomy

Wu and Palmer semantic measure [29] between two concepts or synsets, say, c_1 and c_2 compares the depth of the lowest common subsumer of two concepts to depth of individual concepts as in the following equation.

$$Sim_{wup}(c_1,c_2)=\frac{2*depth(lcs(c_1,c_2))}{depth(c_1)+depth(c_2)} \tag{1}$$

Similarly, Leacock and Chodorow [30] developed a similarity metric that uses the shortest path between the two concepts, normalized using the maximum depth of the taxonomy:

$$Sim_{lch}(c_1,c_2)=-log\frac{len(c_1,c_2)}{2\max\_depth} \tag{2}$$

A third simple measure employing WordNet hierarchy makes use only the path length[1] between the two associated synsets [37]:

[1] One shall also mention the all the three semantic similarity use the concept of path length as well, and are sometimes cited under the category of path length measures.

$$Sim_{path}(c_1, c_2) = 1/len(c_1, c_2) \tag{3}$$

It should be noted that Sim_{wup}, Sim_{path} measures are normalized within the unit interval, while Sim_{lch} ranges from log 2 to log (2max_*depth*). Normalization in unit interval can be achieved using:

$$Sim^{*}_{lch}(c_1, c_2) = \frac{Sim_{lch}(c_1, c_2) - \log 2}{\log((2\max\_depth)) - \log 2} \tag{4}$$

The similarity between two words, say, w_1 and w_2, is generated from the similarity of the synsets they induce as follows:

$$Sim(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} Sim(c_1, c_2), \tag{5}$$

where $s(w_1)$ (resp. $s(w_2)$) is the set of synsets (concepts) in WordNet taxonomy that are senses of word w_1 (resp. w_2).

3.1 Properties of taxonomy-based semantic similarity measure

First, looking at the hyponymy/hypernymy relation *R*: "…IS…A…" or "…IS A KIND OF…", denoted "$@\rightarrow$", which has already been pointed out in Section A, reveals that *R* acts as a partial order relation over the set of all synsets (or concepts). Indeed,

- *R* is trivially reflexive;
- *R* is transitive: for any synsets c_1, c_2, c_3 such that $c_1 @\rightarrow c_2$ $@\rightarrow c_3$, entails $c_1 @\rightarrow c_3$. For example, since "dog" is a *hyponym* of "mammal" and "mammal" is a hyponym of "animal", "dog is a *hyponym* of animal".
- *R* is anti-symmetric: for any synsets c_1, c_2, if $c_1 @\rightarrow c_2$ and $c_2 @\rightarrow c_1$ entails $c_1 = c_2$.

The partial ordering follows from the fact that there are synsets, which are not related by hyponymy relationship.

However, the translation of the hyponym relationship into semantic relations in the sense of the above is not straightforward. A possible question of interest is whether there is any relationship between the value of the semantic similarity and the occurrence or absence of any hyponymy relation. In this respect, intuitively, the following holds.

Proposition 1 Synsets c_i and c_j are linked by hyponymy relation if either $c_i = lcs(c_i, c_j)$ or $c_j = lcs(c_i, c_j)$.

The preceding shows that the information about the lowest common subsumer provides a relevant information regarding the existence of hyponymy[2] relation. Nevertheless, such information is not straightforwardly inferred from semantic similarity measures. Let us first investigate the properties of such semantic relations in terms of range of values assigned to each of them, monotonicity and boundary cases.

Unless stated otherwise, one shall use notation $Sim_x(.,.)$ to stand for any of path, Wu and Palmer, normalized Leacock and Chodorow similarity measures. Consider the relation $Sim_x(c_i, c_j)$ "c_i is semantically related to c_j in the sense of Sim_x", then it holds:

- Reflexivity: $Sim_x(c_i, c_i) = 1$.
- Symmetry: $Sim_x(c_i, c_j) = Sim_x(c_j, c_i)$
- $0 \le Sim_x(c_i, c_j) \le 1$

The above properties were trivial and follow straightforwardly from the definitions of the similarity measures in (1) and (3–4). Other properties of Sim_x are summarized in the following Propositions whose proofs are reported to Appendix of this paper.

Proposition 2 For all synsets c_i, c_j, it holds:

$$\text{(i)} \quad \frac{1}{\max\_depth} \le Sim_{path}(c_i, c_j) \le 1 \tag{6}$$

$$\text{(ii)} \quad \frac{2}{\max\_depth + 1} \le Sim_{wup}(c_i, c_j) \le 1 \tag{7}$$

$$\text{(iii)} \quad 0 \le Sim^{*}_{lch}(c_i, c_j) \le 1 \tag{8}$$

$$Sim_x(c_i, c_j) = 1 \Leftrightarrow c_i = c_j \tag{9}$$

Proposition 3

Proposition 3 *demonstrates that the only case where the semantic similarities take their maximum value is when the underlying pair of words belongs to the same synset.*

Proposition 4 *For all synsets c_i, c_j, c_k, c_l,*

$$\text{(i)} \quad Sim_{path}(c_i, c_j) = Sim_{path}(c_k, c_l) \Leftrightarrow Sim_{lch}(c_i, c_j) = Sim_{lch}(c_k, c_l) \tag{10}$$

$$\text{(ii)} \quad Sim_{path}(c_i, c_j) < Sim_{path}(c_k, c_l) \Leftrightarrow Sim_{lch}(c_i, c_j) < Sim_{lch}(c_k, c_l) \tag{11}$$

[2] We mentioned here and in the whole section hyponymy / hypernymy relation for illustration only as it is the dominant relation in WordNet lexical database, but the reasoning applies equally to meronymy, troponymy relations or any other relation that use hierarchical relation only of WordNet.

To prove the above statements, it is enough to see that Sim_{path} and Sim_{lch} are related to each other through log and linear transformations, and since both logarithmic and linear transformations are strictly monotonic functions, the result follows straightforwardly. Besides, the statements in the core of Proposition 4 are also valid for the normalized Leacock and Chodorow similarity Sim^{*}_{lch}. We shall refer to behaviour induced by expressions (10) and (11) to the *monotonic equivalence* property fulfilled by the path-length and Leacock and Chodorow semantic similarities. However, such a monotonic equivalence property does not hold between Sim_{wup} and any of the two other semantic similarities. To see it, one shall consider the following counter-example.

Example 1 Sim_{path} (process, attribute) = 0.2; Sim_{wup} (process, attribute) = 0.5 Sim_{path} (whole, foode) = 0.1667; Sim_{wup} (whole, foode) = 0.5455.

So it is easy to notice that:

Sim_{path} (process, attribute) > Sim_{path} (whole, foode), while

Sim_{wup} (process, attribute) < Sim_{wup} (whole, foode).

Proposition 5 *For all synsets c_i, c_j it holds*:

$$\text{(i)}\quad Sim_{path}(c_i, c_j) \leq Sim_{wup}(c_i, c_j) \tag{12}$$

$$\text{(ii)}\quad Sim_{wup}(c_i, c_j) \leq Sim^{*}_{lch}(c_i, c_j),\ \text{if } len(c_i, c_j) \leq 2 \quad \text{Otherwise, } Sim_{wup}(c_i, c_j) > Sim^{*}_{lch}(c_i, c_j) \tag{13}$$

Proposition 5 shows that for any pair of synsets that are semantically close in WordNet hierarchy (either being synonyms or one is a direct hyponym of another to ensure the condition $len(c_i, c_j) \leq 2$), the path similarity is the most conservative among the three similarity measures. Otherwise, the Wu and Palmer measure is the less conservative one. This is especially relevant when the order of magnitude of the semantic similarity is deemed important.

Statements in Proposition 5 can also be seen as providing some ordering relationship among the three semantic similarity measures,

Proposition 6

$$c_i \neq c_j \Rightarrow Sim_{path}(c_i, c_j) \leq 0.5 \text{ and } Sim^{*}_{lch}(c_i, c_j) < 0.7 \tag{14}$$

The proof of the statement (14) follows from the fact that in the case of different synsets, then trivially $len(c_i, c_j) \geq 2$, which, after putting the lower bound "2" in (3) and (4) and noticing that the maximum depth in WordNet 3.0 is 20, is translated into inequalities pointed out in the core of this Proposition.

Proposition 6 indicates that Wu and Palmer similarity is the only one that allows the user to expect to obtain high similarity values, close to one when using different synsets. From this perspective, Sim_{wup} has some advantages with respect to the other two semantic similarity measures, especially when thresholding-like approach was employed. In other words, the range of values that can be attributed to Sim_{path} and Sim^{*}_{lch} contain empty slots of (0, 5 1] and (0.7 1], respectively.

Another interesting property concerns the behaviour of these semantic similarity measures when one of the synsets is a direct hyponym of the other one. Strictly speaking, intuitively, a (full) equivalence relation between the hyponymy and semantic similarity relations cannot be held as the former is anti-symmetric while the latter is symmetric. Nevertheless, this does not exclude the existence of hints and/or links between the two concepts. In this course, the following holds.

Proposition 7 Assume synsets $c_1@ \rightarrow c'_1, c_2@ \rightarrow c'_2$, then it holds.

(i) If c_1, c_2, c'_1, c'_2 have the same lower common subsumer, then $Sim_*(c_1, c_2) \leq Sim_*(c'_1, c'_2)$

(ii) If c'_1 and c'_2 are direct hyponyms of c_1 and c_2, respectively, and do not share lower common subsumer, then $Sim_*(c_1, c_2) \geq Sim_*(c'_1, c'_2)$. Especially, $Sim_*(c_1, c_2) \geq Sim_*(c'_1, c'_2)$ for path and Leacock and Chodorow semantic similarities.

(iii) If $c'_1 (resp. c'_2)$ is a direct distinct hyponym of c'_2 (resp.c'_1), then no stable relationship to Sim_* exits.

Proposition 7 indicates that the hyponymy relationship among synsets does not extend straightforwardly to semantic similarity of pairs of synsets. Especially, it is found that the preservation of the monotonicity relationship is guaranteed only when the pairs share the same lowest common subsumer. Otherwise, it has also been pointed out that path and Leacock/Chodorow semantic similarities also conserve the monotonicity in the case of direct hyponymy relationship or when one of the elements of the pair is the lowest common subsumer of the other element.

Especially, in the case of three synsets only, say, c_1, c_2, c_3 where $c_1@ \rightarrow c'_1$, then it holds that.

(i) If c_1, c_2, c'_1 share the same lowest common subsumer, then $Sim_*(c_1, c_2) \leq Sim_*(c'_1, c_2)$.

(ii) Similarly, if $c_2@ \rightarrow c'_2$, where c'_2 share the same lowest common subsumer with (c_1, c_2), then $Sim_*(c_1, c_2) \leq Sim_*(c_1, c'_2)$.

On the other hand, regarding the boundary values of the various semantic similarity measures when one of the synset is a direct hyponym of the other one, the following holds.

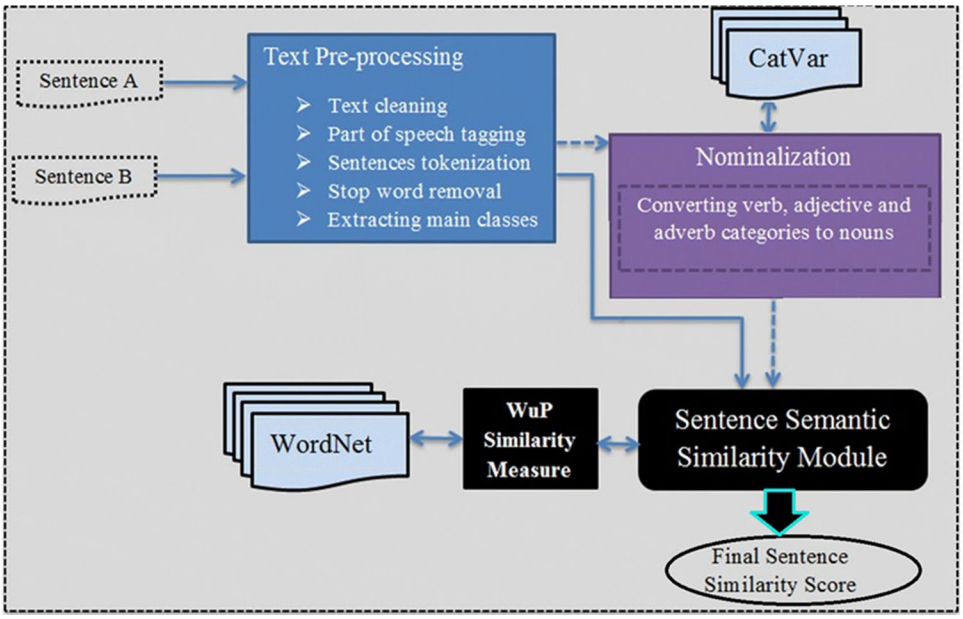

Fig. 2 Illustration of the sentence similarity with all-to-noun conversion

Proposition 8 *Assume that c_i is a direct hyponym (or hypernym) of c_j, then it holds that.*

$$\text{(i)} \quad Sim_{wup}(c_i, c_j) \geq 0.8 \tag{15}$$

$$\text{(ii)} \quad Sim_{path}(c_i, c_j) = 0.5 \tag{16}$$

$$\text{(iii)} \quad Sim^{*}_{lch}(c_i, c_j) = 1 - \frac{\log(2)}{\log(2\max\_depth) - \log(2)} \tag{17}$$

It should be noted that the results pointed out in Proposition 8 are not necessarily held in reverse order; namely, for instance, if $Sim_{wp}(c_i, c_j) = 0.8$, this does not necessarily entail that c_i (resp.c_j) is a hyponym of c_j (resp.c_i).

However, the reverse implication holds in case of path or normalized Leacock and Chodron semantic similarity measures because whenever the length between the two synsets is two, it implicitly entails that one is a direct hyponym of the other one.

Expressions (15–17) provide a formal framework about the incremental evolution of the hyponymy/hypernymy relation.

Proposition 9 Given a sequence of hyponymy relations as

$$c_1@ \rightarrow c_2@ \rightarrow c_3@ \rightarrow \ldots c_{n-1}@ \rightarrow c_n, \tag{18}$$

then it holds that

$$\begin{aligned} &Sim_*(c_1, c_2) \geq Sim_*(c_1, c_3) \\ &\geq Sim_*(c_1, c_4) \geq \ldots \geq Sim_*(c_1, c_{n-1}) \\ &\geq Sim_*(c_1, c_n) \end{aligned} \tag{19}$$

Proposition 9 indicates that when a direct hyponym is available, this yields the highest semantic similarity measure with its associated hypernym among all possible other distinct hyponyms. Proposition 9 also highlights typical scenarios in which the monotonicity of the hyponym relation is preserved when translated into a semantic similarity measure.

The result pointed out in Proposition 9 is in a full agreement with the intuition behind the concept of synset in the sense that the more the hyponym synset is close to the current synset, the more one expects the underlying semantic similarity becomes higher. The preceding reveals the importance of the concept of direct hyponymy to ensure the satisfaction of the monotonicity relation. To see it, it suffices to see the example of Fig. 2a, b in Appendix where c'$_2$ is also a hyponym (but not a direct hyponym) of c_1 and nothing prevents $Sim_*(c_1, c_2)$ to be greater than $Sim_*(c_1, c'_2)$.

Finally, exploring the computational complexity of the three semantic similarity measures, and denoting by $T(Sim_x)$, the execution time of the similarity measure Sim_x, the following holds (in asymptotic behaviour).

Proposition 10 $T(Sim_{path}) \leq T(Sim_{lch}) \leq T(Sim_{wup})$

Table 1 Summary of individual properties of the similarity measures

Similarity measure	Range	Monotonicity with respect to hyponymy/hypernym	Incremental monotonicity evolution	Reflexivity	Symmetry
Path length	$\frac{1}{\max\_depth} \leq Sim_{path}(c_i, c_j) \leq 0.5$ $Sim_{path}(c_i, c_j) = 1$ *if same synsets*	$Sim_{path}(.,x)$ or $Sim_{path}(x,.)$ is monotonic $Sim_{path}(x,y)$ is monotonic if all inputs have same lcs	$Sim_{path}(x,y) = 0.5$ when x is a direct hyponym/hypernym of y	Yes	Yes
Leacock and Chodorow	$0 \leq Sim^{*}_{lch}(c_i, c_j) < 0.7$ $Sim^{*}_{lch}(c_i, c_j) = 1$ *for same synsets*	$Sim_{lch}(.,x)$ or $Sim_{lch}(x,.)$ is monotonic $Sim_{lch}(x,y)$ is monotonic if all inputs have same lcs	$Sim_{lch}(x,y) = 1 - \frac{\log(2)}{\log(2\max\_depth) - \log(2)}$ when x is a direct hyponym/hypernym of y	Yes	Yes
Wu and Palmer	$\frac{2}{\max\_depth+1} \leq Sim_{wup}(c_i, c_j) \leq 1$	$Sim_{wup}(.,x)$ or $Sim_{wup}(x,.)$ is monotonic $Sim_{wup}(x,y)$ is monotonic if all inputs have same lcs	$Sim_{wup}(x,y) \geq 0.8$ when x is a direct hyponym/hypernym of y	Yes	Yes

To prove the above inequality, consider two concepts c_1, c_2 (same reasoning applies for two words w_1, w_2 as well by maximization over all induced concepts), one shall notice the following.

The calculus of $len(c_1,c_2)$ required in Sim_{path} and Sim_{lch} can be performed in $O(V+E)$ time where V and E stand for the number of vertices and edges in the WordNet hierarchy (using Breadth-first search algorithm).

Noticing that the least common subsumer $lcs(c_1,c_2)$ can be inferred straightforwardly when a shortest path from c_1 and c_2 ($len(c_1,c_2)$) is found without any additional exploration of the network, while additional network scrutinizing is required to calculate $depth(c_1)$ and $depth(c_2)$ in Sim_{wup}. This yields $Sim_{path}(c_1, c_2) \leq Sim_{wup}(c_1, c_2)$ and $Sim_{lch}(c_1, c_2) \leq Sim_{wup}(c_1, c_2)$.

To compare $T(Sim_{path})$ and $T(Sim_{lch})$, the issue boils down to the comparison between the computational complexity of the arithmetic division and logarithmic operations. In this respect, if $len(c_1,c_2)$ and the outcome of the arithmetic/logarithmic operations are n-digit number, then the arithmetic division operation in Sim_{path} can be executed in $O(M(n))$ where M(n) stands for the complexity of the chosen multiplication, using, for instance, Newton–Raphson division algorithm, while the complexity of the logarithmic operation $O(M(n)k)$ for some constant k. Therefore, $Sim_{path} \leq Sim_{lch}$ on asymptotic level. See Borwein [42] for further details on complexity on arithmetic and functional operations.

3.2 Summary

Table 1 summarizes the main properties of the three semantic similarities: path-length, Wu and Palmer, Leacock and Chodorow. We shall distinguish between properties fulfilled by a single semantic similarity irrespective of others and properties that highlight the interactions among the various semantic similarities. Individual properties of similarity measures include reflexivity, symmetry, range, (one argument, two-arguments) monotonicity with respect to hyponymy/hypernymy relationship and incremental monotonicity evolution.

In Table 2, we summarize the interactive properties of the three semantic similarities highlighting the properties of score ordering, time complexity and monotonic equivalence.

These properties can provide valuable insights to any researcher or computational linguistic specialist who is interested in applying WordNet-based semantic similarity in text mining and retrieval as will be detailed in the discussion section of this paper. This also yields another look and revisit of the results that might be expected from WordNet-based semantic similarity analysis.

4 Sentence semantic similarity

4.1 From word semantic similarity to sentence similarity

Since the sentence is constituted of a set of words, the sentence-to-sentence semantic similarity is intuitively linked to word-to-word semantic similarities, although a sentence is more than a simple bag of words because of the importance of word disposition, part of speech, punctuation, among

Table 2 Summary of interactive properties of the similarity measures

Property	Fulfilment
Score ordering	$Sim_{path}(x,y) \leq Sim_{wup}(x,y)$ for all x, y $Sim_{wup}(x,y) \leq Sim^{*}_{lch}(x,y)$, if $len(x,y) \leq 2$, Otherwise $Sim_{wup}(x,y) > Sim^{*}_{lch}(x,y)$
Time complexity	$T(Sim_{path}) \leq T(Sim_{lch}) \leq T(Sim_{wup})$
Monotonic equivalence	Sim_{path} and Sim_{lch} are monotonically equivalent

others, in conveying a specific meaning to the sentence. For this purpose, various linguistic manipulations were suggested to combine word semantic values to achieve sentence-to-sentence semantic scores. Borrowed from information retrieval techniques, traditionally, words of each sentence are represented using weights built from some corpus-like information, e.g., information content of each word, frequency times inverse document frequency (TF-IDF), among others [3, 13, 14], then one uses a cosine-like similarity measure to infer the overall score of the similarity between the two input sentences. Nevertheless, such reasoning fails to account for the WordNet semantic similarity measure. Li et al. [25, 26] proposed a hybrid approach that combines both vector-based and WordNet models, where the overall sentence similarity score is given as a convex combination of semantic similarity and word-order similarity. The latter provides a rough indication of how similar the word order between the two sentences is. The suggested sentence semantic similarity accounts for both information content and path-length semantic similarity. A simple extension of the word semantic similarity to sentence similarity that accounts for word part of speech but discarding the explicit corpus-based information is suggested by Mihalcea et al. [23] where the similarity between the pair of words that ensures the maximum score among all other pairs is accounted for. This approach has also been used in the query system [40]. It is worth noticing that the computation of word-pairwise similarity scores, using either path-based or information content, can only be enabled if the terms belong to the same part of speech. More formally, expression (20) forms the basis of the sentence semantic resemblance between sentences S_A and S_B [23]:

$$Sim_g^*(S_A, S_B) = \frac{\sum_{w\in S_A} \max_{\substack{x\in S_B,\\ \text{PoS(x)=PoS(w)}}} Sim_*(w,x) + \sum_{w\in S_B} \max_{\substack{x\in S_A,\\ \text{PoS(x)=PoS(w)}}} Sim_*(w,x)}{|S_A| + |S_B|} \tag{20}$$

where the semantic similarity of the word w of sentence S_A and sentence S_B, $Sim_*(w, S_B)$, is given as the semantic similarity of the word in S_B of the same part of speech as w yielding the highest semantic similarity with w. Similar reasoning applies to $Sim_*(w, S_A)$.

From (20), it is worth noticing that only verbs and nouns are accounted for in the sentence-to-sentence similarity measure, which suggests that all other part-of-speech wording of the sentence (s) is ignored. The normalization factor in the denominator of expression (20) ensures the overall sentence similarity to be within the unit interval. Trivially, any of word semantic similarity highlighted in expressions (1–3) can be employed to quantify $Sim_*(w, x)$. A more concise formulating of (20) that accounts for normalization issues and distinguishes the contribution of both sentences to overall semantic similarity, in the same spirit as Mihalcea et al.'s proposal [23],[3] can be provided as.

$$Sim_g^*(S_A, S_B) = \frac{1}{2}\left(\frac{\sum_{w\in S_A} \max_{\substack{x\in S_B,\\ \text{PoS(x)=PoS(w)}}} Sim_*(w,x)}{|S_A|} + \frac{\sum_{w\in S_B} \max_{\substack{x\in S_A,\\ \text{PoS(x)=PoS(w)}}} Sim_*(w,x)}{|S_B|} \right) \tag{21}$$

Because of use of $Sim_*(,)$, expression (21) assumes that the semantic scores are only linked to *WordNet* taxonomy without any reference to the corpus. In order to exemplify expression (21), one can write sentences S_A and S_B as bag of nouns and verbs (other part of speech, e.g. adverbs, adjectives as well as any stopwords do not contribute to (21)):

S_A: $N_{11}, N_{21}, ..N_{p1}, V_{11}, V_{21}, \ldots, V_{q1}$

S_B: $N_{12}, N_{22}, ..N_{m2}, V_{12}, V_{22}, \ldots, V_{n2}$where p, m are the number of nouns in sentences A and B, respectively, while q and n stand for the number of verbs in sentences A and B, respectively. Let us consider permutations σ and ϕ defined on sets {1,2,…, *max(p,m)*} and {1,2,…,*max(q,n)*)}, respectively, such that:

$$\begin{aligned} S_A &: N_{\sigma(1)1}, N_{\sigma(2)1}, \ldots, \\ &N_{\sigma(max(p,m))1}, V_{\phi(1)1}, V_{\phi(2)1}, \ldots, \\ &V_{\phi(max(q,n))1}\ S_B : N_{\sigma(1)2}, N_{\sigma(2)2}, \ldots, \\ &N_{\sigma(max(p,m))2}, V_{\phi(1)2}, V_{\phi(2)2}, \ldots, \\ &V_{\phi(max(q,n))2} \end{aligned}$$

With

$$\max_{i=1,m} Sim_*(N_{j1}, N_{i2}) = Sim_*(N_{j1}, N_{\sigma(i)2}),\ \text{j} = 1 \text{ to } p \tag{22}$$

$$\max_{i=1,n} Sim_*(V_{j1}, V_{i2}) = Sim_*(V_{j1}, V_{\phi(i)2}),\ \text{j} = 1 \text{ to } q, \tag{23}$$

$$\max_{i=1,p} Sim_*(N_{j2}, N_{i1}) = Sim_*(N_{j2}, N_{\sigma(i)1}),\ \text{j} = 1 \text{ to } m \tag{24}$$

$$\max_{i=1,q} Sim_*(V_{j2}, V_{i1}) = Sim_*(V_{j2}, V_{\phi(i)1}),\ \text{j} = 1 \text{ to } q, \tag{25}$$

It should be noted that extra nouns and/or verbs in sentences S_A or S_B correspond to a duplication of already existing nouns/verbs in order to ensure that that the pair ($N_{\sigma(i)1}$ $N_{\sigma(i)2}$) yields the highest semantic similarity measure. Similar reasoning applies to pairs ($V_{\phi(i)1}$ $V_{\phi(i)2}$) ($i = 1,2\ldots$). As

[3] The notations employed were slightly different from the original expression of [23].

a consequence of the above construction, expression (21) boils down to

$$Sim_g^*(S_A, S_B) = \frac{\sum_{i=1,p} Sim_*(\mathrm{N}_{i1}, \mathrm{N}_{\sigma(i)2}) + \sum_{i=1,q} Sim_*(\mathrm{V}_{i1}, \mathrm{V}_{\phi(i)2})}{2(p+q)} + \frac{\sum_{i=1,m} Sim_*(\mathrm{N}_{i2}, \mathrm{N}_{\sigma(i)1}) + \sum_{i=1,n} Sim_*(\mathrm{V}_{i2}, \mathrm{V}_{\phi(i)1})}{2(m+n)} \quad (26)$$

Example 2 Consider the following set of sentences:

S_A: Students are heavily involved in exams these days.

S_B: Only few students are expected to achieve first class marks.

In the above sentences, tokens "are", "these", "only", "few", "first" are commonly treated as part of stopword list, which are eliminated as part of preprocessing stage associated with the above sentences. Similarly, some tokens are converted into their root forms, e.g. "students" to "student", "days" to "day", "involved" to "involve", etc., as part of preprocessing stage.

Therefore, the use of expression (21), or equivalently (26), yields

$$Sim_g^*(S_A, S_B) = \frac{1}{2}\left(\begin{array}{l} \frac{\max(Sim(student, student), Sim(student, class), Sim(student, mark))}{|S_A|} \\ + \frac{\max(Sim(exam, student), Sim(exam, class), Sim(exam, mark))}{|S_A|} \\ + \frac{\max(Sim(day, student), Sim(day, class), Sim(day, mark))}{|S_A|} \\ + \frac{\max(Sim(involve, s\exp ect), Sim(involve, achieve))}{|S_A|} \\ + \frac{\max(Sim(student, student), Sim(student, exam), Sim(student, day))}{|S_B|} \\ + \frac{\max(Sim(class, student), Sim(class, exam), Sim(class, day))}{|S_B|} \\ + \frac{\max(Sim(mark, student), Sim(mark, exam), Sim(mark, day))}{|S_B|} \\ + \frac{Sim(\exp ect, involve) + Sim(achieve, involve)}{|S_B|} \end{array}\right)$$

– Using Wu and Palmer similarity measure, we have:

$$Sim_g^{WP}(S_A, S_B) = \frac{1}{2}\left(\begin{array}{l} \frac{1+0.71+0.67+0.5}{4} + \\ \frac{1+0.71+0.71+0.5+0.4}{5} \end{array}\right) \approx 0.69$$

– Using path-length measure:

$$Sim_g^{PL}(S_A, S_B) = \frac{1}{2}\left(\begin{array}{l} \frac{1+0.17+0.17+0.33}{4} + \\ \frac{1+0.17+0.2+0.33+0.25}{5} \end{array}\right) \approx 0.40$$

Using (normalized) Leacock and Chodorow similarity measure yields:

$$Sim_g^{NLC}(S_A, S_B) = \frac{1}{2}\left(\begin{array}{l} \frac{1+0.40+0.40+0.58}{4} + \\ \frac{1+0.40+0.46+0.58+0.47}{5} \end{array}\right) \approx 0.59$$

From the above example, it is worth noticing the following:

– The stopwords as well as adjectives/adverbs, although they are important in conveying meaning to the underlined sentence, are not taken into account in the above sentence-to-sentence similarity. An intuitive way to account for the occurrence of such tokens consists of expanding the range of $|S_A|$, for instance, to include all tokens in sentence A including stopwords and adverbs/adjectives. However, this would substantially reduce the score of the sentence-to-sentence similarity. Besides, such integration of extra tokens would not take into account any meaning of such wording so that if another sentence, say S'_A contains adverb/adjective (s), which are antonyms in sentence S_A, still will induce the same scoring value! Consequently, discarding such token seems a rational attitude in this respect.
– Some tokens, like "involved", can appear in both verb and adjective categories. However, since only verbal

entities are present in the taxonomy, the handling of the token as such seems convenient and intuitive.

On the other hand, from the generic expression (21), one can straightforwardly notice some interesting special cases:

- Assume that sentence S_A reduces, after some text preprocessing task, to one single noun N_1 and one single verb V_1, while sentence S_B reduces to noun N_2 and verb V_2, then

$$Sim_g^*(S_A, S_B) = \frac{1}{2}\left(Sim_*(N_1, N_2) + Sim_*(V_1, V_2)\right) \quad (27)$$

- In case of identical sentences, or at least, identical nouns and verbs in both sentences S_A and S_B, then it is easy to see that $Sim_g^*(S_A, S_B) = 1$
- In case where sentences, after preprocessing, reduce to a single noun or verb, then sentence similarity boils down to the corresponding word semantic similarity; that is, assuming N_1 and N_2 (resp. V_1 and V_2) be the nouns (resp. verbs) associated with sentence S_A and S_B, respectively, then

$$Sim_g^*(S_A, B) = Sim_*(N_1, N_2)$$

$$\left(\text{resp. } Sim_g(S_A, S_B) = Sim_*(V_1, V_2)\right)$$

- There are situations in which the semantic similarity between the two sentences is not defined. Indeed, this occurs if the two sentences contain neither naming nor verbal expressions, or in case where one sentence contains only naming expression and the other one only verbal expression. In such cases, there is no analogy between the parts of speech in the two sentences, which renders the application of expression (21) void. Similarly, this also occurs if at least one of the two sentences contains neither verbal nor naming expression (s). An example of such sentences is: S_A: "How are you?"; S_B: "Hi there".
- An interesting case is related to the situation where a sentence contains repeated words or semantically equivalent name or verb expressions. It will thereby be of interest to see how such repetition influences the sentence similarity score. In this course, the following holds.

Proposition 11 Consider two sentences S_1 and S_2 such that.

$$S_1 = \left(N_{11}, N_{12}, ..N_{1p}, V_{11}, V_{12}, ..V_{1q}\right)$$

$$S_2 = \left(N_{21}, N_{22}, ..N_{2s}, V_{21}, V_{22}, ..V_{1l}\right)$$

where p, q (resp. s and l) are the number of names and verbs in sentence S_1 (resp. S_2).

Assume nouns N_{1i}, i = 1 to p (resp. N_{2j}, j = 1 to s) are semantically equivalent. Similarly, verbs V_{1i}, i = 1 to q (resp. V_{2j}, j = 1 to l) are also assumed semantically equivalent. Then

$$Sim_g^*(S_1, S_2) = Sim_g^*(S_1', S_2') \Leftrightarrow \frac{p}{q} = \frac{l}{s} \quad (28)$$

where $S_1' = \left(N_{11}, V_{11}\right)$ and $S_2' = \left(N_{21}, V21\right)$.

Proposition 11 states the conditions under which the use of semantically equivalent nouns and verbs in the sentences preserves the overall semantic similarity score of the original sentences (without introducing extra equivalent nouns and verbs). Especially it has been pointed out that such equivalence holds only if the ratio between the number of equivalent nouns and the number of equivalent verbs in the first sentence is inversely proportional to the ratio yield by the second sentence. This includes, for instance, the case where the two sentences have the same number of equivalent nouns as well as the same number of equivalent verbs. However, if the condition stated in the core of Proposition 11 is not held then the use of semantically equivalent nouns and verbs may either increase or decrease the overall semantic similarity of the two sentences when compared to semantic similarity of original sentences (which do not contain equivalent nouns/verbs). To see it, consider for instance the following sentences:

$$S_1 = \left(N_{11}, N_{12}, V_{11}\right)$$

$$S_2 = \left(N_{21}, V_{21}\right)$$

After some manipulations, one obtains

$$\begin{aligned} Sim_g^*(S_1, S_2) &= \frac{1}{2}\left[\frac{2Sim_*(N_{11}, N_{21}) + Sim_*(V_{11}, V_{21})}{3} + \frac{Sim_*(N_{11}, N_{21}) + Sim_*(V_{11}, V_{21})}{2}\right] \\ &= \frac{1}{2}\left[\frac{7Sim_*(N_{11}, N_{21}) + 5Sim_*(V_{11}, V_{21})}{6}\right] \end{aligned} \quad (29)$$

Consider now the corresponding sentences without additional equivalent verb/noun; namely, $S_1' = \left(N_{11}, V_{11}\right)$ and $S_2' = S_2$. Therefore, $Sim_g^*(S_1, S_2) \geq Sim_g^*(S_1', S_2')$ is equivalent to

$$\left(\frac{7Sim_*(N_{11},N_{21})+5Sim_*(V_{11},V_{21})}{6}\right) \geq \left(\frac{Sim_*(N_{11},N_{21})+Sim_*(V_{11},V_{21})}{2}\right) \tag{30}$$

This entails

$$Sim_*(N_{11},N_{21}) - Sim_*(V_{11},V_{21}) \tag{31}$$

The latter inequality is not systematically held and is rather context dependent. To see it, let us consider the following example.

Example 4 S_1: These persons look great among available individuals → (person, individual, look).

S_2: Students seem active (student, seem)

$$Sim_g^*(S_1,S_2) - Sim_g(S_1',S_2') = Sim_*(individual,student) - Sim_*(look,seem) = 0.33 - 1 < 0$$

(using path semantic similarity).

Proposition 12 Consider two sentences S_1 and S_2 such that

$$S_1 = (N_{11},N_{12},..N_{1p},V_{11},V_{12},..V_{1q})$$

$$S_2 = (N_{21},N_{22},..N_{2p},V_{21},V_{22},..V_{2q})$$

Assume there is one-to-one direct hyponymy relation:

$$N_{1i}@ \rightarrow N_{2i};\ V_{1j}@ \rightarrow V_{2j},\ i=1,p; j=1,q.$$

Then, it holds

$$Sim_g^*(S_A,S_B) = \frac{\sum_{i=1}^{p} Sim_*(N_{1i},N_{2i}) + \sum_j^q Sim_*(V_{1i},V_{2j})}{p+q} \tag{32}$$

Proof Expression (32) follows straightforwardly from Proposition 10 and application of (21) in case of sentences S_A and S_B having the same number of verbs and nouns. The detail is omitted.

Strictly speaking, Proposition 12 provides an example of how the hyponymy relation can be used to simplify expression (21). Interestingly, the outcome pointed out in Proposition 12, as highlighted in expression (32), is similar to that of (26) when sentences S_A and S_B have the same number of nouns and verbs, where permutations boils down to identity, e.g. $\sigma(i) = \varnothing(i) = i$ for $i = 1,2,..$

Another question of interest relates to the behaviour of the sentence similarity rule with respect to a possible contraction or expansion, namely given sentences S_A and S_B, where

S_A: N_{11}, N_{21},..N_{p1}, V_{11},V_{21},…,V_{q1}.

S_B: N_{12}, N_{22},..N_{m2}, V_{12},V_{22},…,$V_{n2,}$

Then given a new sentence S'_A (resp. S'_B) issued from S_A (resp. S_B) by, say, omitting some of its terms, e.g.

S'_A: N_{11}, N_{21},..$N_{p'1}$, V_{11},V_{21},…, $V_{q'1}$, $p' < p$, $q' < q$.

S'_B: N_{12}, N_{22},..$N_{m'2}$, V_{12},V_{22},…,$V_{n'2,}$ $m' < m$, $n' < n$,the problem here is how the semantic similarity of sentence S_A and S_B compares to that of sentences S'_A and S'_B? Generic answer to such question by either increase or decrease in sentence semantic similarity without any further constraint does not exist. Nevertheless, the following provides a skeleton of an answer to such problem.

Proposition 13 Let

S_A: N_{11}, N_{21},..N_{p1}, V_{11},V_{21},…,V_{q1}.

S_B: N'_{12}, N'_{22},..N'_{p2}, V'_{12},V'_{22},…,$V'_{q2,}$ such that

$\max_{i=1,p} Sim_*(N_{j1},N'_{i2}) = Sim_*(N_{j1},N'_{j2})$ j = 1 to p, and

$$\max_{i=1,p} Sim_*(N_{j1},N'_{i2}) = Sim_*(N_{j1},N'_{j2})$$

If a new word, say noun $N_{p+1,2}$ is added to sentence S_B (similar result applies to $V_{q+1,2}$), yielding a new sentence.

S'_B: N'_{12}, N'_{22}, N'_{p2}, $N_{p+1,2.}$ V'_{12}, V'_{22},…,$V'_{q2,}$

Then $Sim_g^*(S_A,S_B) \leq Sim_g^*(S_A,S'_B)$ if and only if there exists a noun N_{k1} in sentence S_A such that

$$\max\left(\begin{matrix}\max\limits_{i=1,q} Sim_*(V_{i1},V'_{i2}),\ \max\limits_{i=1,p} Sim_*(N_{i1},N'_{i2}),\\ \max\limits_{i=1,p} Sim_*(N_{i1},N_{p+1,2})\end{matrix}\right) = Sim_*(N_{k1},N_{p+1,2}) \tag{33}$$

Proposition 13 indicates that adding an extra element to the sentence would bring an increase in semantic sentence similarity if the added element is sufficiently semantically similar to one of the elements of the sentence in the sense of yielding the highest semantic similarity score among all possible pairs between the elements of the two sentences. As a special case, if one adds an element which is a synonym to one of the components of the sentence, then this always yields an increase of the semantic similarity. This is because the use of synonyms would make the quantity $Sim_*(N_{k1},N_{p+1,2})$ equal to one, which renders constraint (33) in Proposition 13 always valid.

Proposition 13 can easily be extended in case of adding more than one element as follows.

Proposition 14 Using same arguments of Proposition 13, given.

S''_B: N'_{12}, N'_{22},..N'_{p2}, $N_{p+1,2.,}$ $N_{p+s,2..,}$ V'_{12},V'_{22},…,$V'_{q2,}$

Then

$Sim_g^*(S_A, S_B) \leq Sim_g^*(S_A, S_B'')$ if and only if there exist nouns N_{k_i}1(i = 1 to s) in sentence S_A such that

$$\left\{ \begin{array}{l} Sim_*\left(N_{k_1 1}, N_{p+1,2}\right) \\ = max(\max\limits_{i=1,q} Sim_*\left(V_{i1}, V'_{i2}\right), \max\limits_{i=1,p} S\left(N_{i1}, N'_{i2}\right), \max\limits_{i=1,p} Sim_*\left(N_{i1}, N_{p+1,2}\right)) \end{array} \right\}$$

$$\left\{ \begin{array}{l} Sim_*\left(N_{k_1 1}, N_{p+s,2}\right) \\ = max(\max\limits_{i=1,q} Sim_*\left(V_{i1}, V'_{i2}\right), \max\limits_{i=1,p} Sim\left(N_{i1}, N'_{i2}\right), \max\limits_{i=1,p} Sim_*\left(N_{i1}, N_{p+s,2}\right)) \end{array} \right\}$$

The proof of Proposition 14 follows the same spirit as that of Proposition 13, so the detail is omitted.

Proposition 14 shows that adding extra elements to the original sentence increase the overall sentence similarity score as long as each of the added elements is sufficiently similar to one of the elements of the original sentence in the sense described above. More generally, this can be written

$$S_B \subset S'_B \Rightarrow Sim_g^*\left(S_A, S_B\right) \leq Sim_g^*\left(S_A, S'_B\right) \tag{34}$$

under constraint (33), for any non-empty (in terms of both noun and verb expressions) sentence S_A.

The inclusion operator employed in (34) is in the sense of set-inclusion of bag of words constituted by the set of nouns and verbs of each sentence.

On the other hand, it should be noted that the disposition of the sentences S_A and S_B described in the core of Proposition 13 is not a prerequisite for the result as this can be deduced from any pair of sentences using appropriate permutation of set of indices as already pointed out in (22–26).

In summary, one shall mention in response to research question Q3 that except trivial symmetry and reflexivity relations, other word-to-word-level similarity properties are not straightforwardly extended to sentence-to-sentence similarity. For instance, score ordering and time complexity are fulfilled in the sense that $Sim_{path}(S_A, S_B) \leq Sim_{wup}(S_A, S_B)$ and $T(Sim_{path}(S_A, S_B)) \leq T(Sim_{lch}(S_A, S_B))) \leq T(Sim_{wup}(S_A, S_B))$.

However, further constraints are still required to meet the monotonicity property for instance.

5 Effect of part-of-speech conversion

So far, the handling of nouns and verbs part of speech were performed independently in (21). However, such independence can also be debatable from different perspectives. First, the dissymmetry of verbs and nouns in WordNet taxonomy cannot be ignored as already pointed out in the introduction of this paper at least with respect to the parameters of the hierarchy since the number of noun synsets, maximum depth and average number of hyponyms are substantially higher in case of nouns [2]. Second, nouns and verbs that have the same lexical root are very common in the English language. For instance, the words "connect" and "connection" have the same semantic meaning. Third, it is quite common that a single word does support several part-of-speech categories in the sense that it can be considered as a noun as well as a verb according to context. Fourth, there are often named-based or verbal-based sentences where either verb or noun category is missing, which renders the application of quantification (21) almost void. To see it, let us consider the following example.

Example 5 S_A: Housing is included in the tuition fee.
S_B: The hostel is part of the tuition fee.

After, an initial text preprocessing, the first sentence boils down to (Housing, tuition, fee, include) while sentence B entails tokens (hostel, part, tuition, fee). It is therefore easy to notice that unlike sentence A, sentence B contains no verb category, which would make the word "include" not contributing to the overall similarity sentence score.

To handle this difficulty, the key is to use the existing linguistic transformations which can turn verb, adjective, adverb into the corresponding noun and vice versa, see [41, 43–45] for details and exemplifications. This motivates this subsequent section where verb–noun conversations are investigated to enhance the sentence semantic similarity calculus. Especially, it is expected that such conversion would allow us to overcome the possible ambiguity or inconsistency where (21) accounts equally for both verb and noun pairwise semantic similarity measures. Besides, it is also expected to account for both adjective and adverb categories in the semantic similarity calculus. For instance, in Example 5, if a such conversion took place, "include" would be converted into, say, noun "inclusion", which provides a good counterpart to word "part" in sentence B when computing word semantic similarity. As such, the transformation of this verb to the same part of speech as the other constituent words in the other sentence would raise the similarity score using (21) from 0.89 to 0.92. The effect of this single verb–noun conversion is therefore revealed in the change of the similarity score, where, intuitively, as the two sentences have close semantic meaning, one expects high semantic similarity scores.

Table 3 Comparison of the sentence similarity scores using the different methods and correlation with human judgment

Methods	Correlation r_s	Mean value of similarity score of sentence pairs	Min–Max value of similarity	Standard deviation	Median	Processing time (sec)
STASIS	0.816	0.589433	[0.209 1]	0.193619	0.6145	0.561
LSA	0.838	0.687667	[0.505 1]	0.143315	0.685	0.134
WN	0.821	0.656186	[0.362 1]	0.168924	0.6272	0.343
WNwC	0.846	0.695833	[0.397 1]	0.155162	0.683	0.423

On the other hand, one can also think of using a noun-to-verb conversation. In this case, the new calculus of the sentence similarity score decreases to 0.402. The reason for this is that a such strategy cannot handle some of the nouns e.g. "university", "tuition", "year" (because of the absence of verbal counterpart in WordNet taxonomy), which, in turn, will negatively influence the overall score of the sentence similarity. Therefore, given the more elaborated structure of noun category and the existence of semantically related nouns for most verbs, adjectives and adverbs, the transformation all-to-nouns is more appealing. For this purpose, an approach employing WordNet taxonomy has been put forward and investigated in our previous work reported in [38, 39, 41], although further work is still required in order to explore the intuitive and physical constraints induced by such transformations.

The high-level description of the implementation of this transformed sentence-to-sentence similarity is illustrated in Fig. 2.

In order to evaluate the performance of the all-to-noun transformation on semantic similarity score, we have conducted an experiment using Benchmark dataset and with comparison with some state-of-the-art methods.

6 Dataset

We used O'Shea et al. [46] dataset, which consists of 65 sentence pairs where human similarity judgements are assigned to each pair. During this dataset creation, 32 graduate native speakers were assigned to score the similarity degree between each pair using scores from 0.0 to 4.0 so that a score 0 corresponds to unrelated sentence pair, 1 for vaguely similar in meaning, 2 for sentences very much a like in meaning, 3 for sentences strongly related in meaning and 4 for sentences that are identical in meaning.

7 Evaluation of semantic similarity measures

Given the provided human judgement in terms of the similarity between the sentences of each pair, we can evaluate the usefulness of the elaborated semantic similarity Sim by calculating the correlation coefficient between the similarity measure and human judgment using the following expression, where $h_i, Sim(p_i)$ stand for the human and machine (via semantic similarity Sim) assigned score to the ith pair of sentences and n is the total number of pairs:

$$r_s = \frac{n\sum_{i=1}^{n} h_i Sim(p_i) - \sum_{i=1}^{n} h_i \sum_{i=1}^{n} Sim(p_i)}{\sqrt{\left(n\sum_{i=1}^{n} h_i^2 - \left(\sum_{i=1}^{n} h_i\right)^2\right)}\sqrt{\left(n\sum_{i=1}^{n} \left[Sim(p_i)\right]^2 - \left(\sum_{i=1}^{n} \left[Sim(p_i)\right]\right)^2\right)}}$$

When another semantic similarity measure ($Sim(p_i)$) is employed another correlation index r_s is generated. We shall consider two state-of-the-art and baseline models in this experiment in addition to the canonical extension (26).

- STASIS corresponds the case where the semantic similarity is obtained using Lie et al. [25, 26] sentence similarity based on semantic net and corpus statistics.
- LSA corresponds to the case where the sentence similarity obtained using latent semantic analysis approach [47].
- WN corresponds to the case where the sentence similarity is calculated using the canonical extension (26) with Wu and Palmer word-to-word semantic similarity measure. The choice of Wu and Palmer similarity is justified by the behaviour of the similarity measures, especially in terms of the range of the values that can be assigned to. In the implementation, we used Pederson's [48, 49] implementation module of the Wu and Palmer semantic similarity measure. For preprocessing, we used Illinois Part-of-speech tagger [50] to identify various test segments.
- WNwC corresponds to the case where the semantic similarity is calculated using the canonical extension (26) with Wu and Palmer word-to-word semantic similarity measure after performing the "all-to-noun" conversation. This conversation is performed using the

Table 4 Statistical Significant Test (T-test) results at a 95% Confidence Interval

Paired variables	MoDs	Lower interval	Upper interval	T value	*P* value
STASIS–WNwC	−0.1064	-Inf	−0.0597025	−3.8715	0.0002833
WN–WNwC	−0.0396474	-Inf	−0.01611506	−2.8627	0.003861
LSA–WNwC	−0.0081667	-Inf	0.02323585	−0.44188	0.3309

CatVar model [43] because it is found to yield better results than morphosemantic links [44, 45].

We computed the sentence-to-sentence similarity of each pair of the dataset using STASIS, LSA, WN and WNwC methods. We next calculated the correlation index with human judgments for each of the above methods. Finally, to gather the statistics about each method, we also calculated the mean, median, standard deviation, maximum and minimum values of the similarity score over all pairs of the dataset. The result of this analysis is summarized in Table 3.

The results summary in the table shows that the WNwC (WordNet-based all-to-noun conversion) method achieves better performance than other benchmark methods in terms of correlation with human judgment, although we noticed that the performance of the LSA algorithm is quite close with less variability in the scores than our all-to-nouns approach).

Table 3 also reports indication about the execution time of each method when performed on a standard PC machine equipped with Intel Core i5-10400F processor and 16 GB RAM. The results show that the developed approach, although induces extra processing time to handle the PoS conversion task as compared to conventional WN model, the execution time can fit soft real-time requirement, and it outperform the state-of-the-art STASIS model, which involves an extra component associated with word ordering similarity that may be computationally burdened. It should be noted that the above results were obtained using our local re-implementation of the STASIS that uses locally stored WordNet semantic similarity values. Otherwise, if one uses the WordNet server to compute the semantic similarity, the execution time will be much longer and the result will be less reliable as it would depend on the server availability and traffic as well. However, LSA is computationally much more effective because it relies on solely on matrix manipulation and no external lexical database.

Next, in order to find out whether the results in terms of correlation with human judgment are statistically significant, we compared the population of the semantic similarity scores of various pairs using each method and used a statistical t-test.

More formally, we have run a t-test comparing the sentence similarity scores of the 65 sentence pairs and analysed the difference of each method with that of WNwC to see whether there is a statistical difference between the two approaches. Through the statistical tests, as highlighted in Table 4, we found that the WordNet-based all-to-noun conversion significantly improves over the conventional sentence semantic similarity as it yields ($t(64) = 3.5986$, $p < 0.0005$). This means that we have evidence to reject the Null hypothesis and conclude that the improvement with the all-to-noun conversion is statistically significant.

The paired t test results in Table 4 also confirmed that WordNet-based PoS conversion method achieves significant improvement over the STASIS ($p < 0.0005$) similarity measure. However, the improvement over LSA method does not seem significant, although the associated t value (0.44188) lies in the required confidence interval (-Inf—0.02323585).

8 Discussions and implications

The results highlighted in this paper provide significant insights and implication from theoretical and practical perspective.

- The algebraical and interactive properties of the investigated WordNet semantic similarities, namely path-length, Wu and Palmer and Leacock and Chodorow measures, provide valuable insights to data mining or natural language processing researchers and practitioners.

  To exemplify this reasoning, consider a commonly employed example of designing a data mining task that uses a thresholding on a semantic similarity measure, say, $Sim_x(c_1, c_2) \geq \zeta$, where, often the threshold ζ is chosen either empirically or by imposing some default values. Nevertheless, if a path-length measure was employed, and we set any value$\zeta > 0.5$, would yield a useless outcome. This is because we can predict through the results pointed out in Sect. 2 that such threshold will only enable us to capture one single instance corresponding to synonyms. The same reasoning applies for Leacock and Chodorow measure when we set a threshold $\zeta > 0.7$. However, such restriction does not apply in case of Wu and Palmer measure.

  Similarly, the incremental evolution property provides us with useful insights in terms of what we may expect as a result when one slightly changes the input along the lexical hierarchy. Especially, it shows that an incremental moves up or down within the hierarchy as

in direct hyponymy/hypernymy relation would yield a constant similarity score in case of path-length or Leacock and Chodorow measures, regardless of the words employed, while Wu and Palmer measure ensures a high score (beyond 0.8) whose exact value depends on the individual words employed. Such knowledge can be very useful, for instance, to predict the robustness of a given plagiarism detection system that is built on WordNet semantic similarity layer. Likewise, the interactive properties show, for instance, that if hard time requirements were imposed, then path-length similarity should be prioritized.

- The results gained from Sect. 3 have direct implications on the type of handling that may be needed for the underlined text mining task. This involves the type of preprocessing that should be performed prior to calling upon the sentence-to-sentence similarity module. This depends on whether the NLP task favours maximizing similarity score or minimizing it, according, for instance, to the criticality of the false positive in the underscored task. This can provide insights whether short sentences are deemed more important or not. And, if rephrasing is permitted, how such operation could be performed in such a way to maximize or minimize similarity score. Especially, the finding highlights the importance of PoS tagging as an initial scrutinizing task. If such analysis reveals that the two input sentences contain only distinct PoS word categories, then one concludes immediately that the sentence-to-sentence similarity yields zero similarity score without any further processing. This also highlights the importance of accurate part-of-speech tagger in the initial scrutinizing stage as any error has a substantial consequence on the outcome. Interestingly, this provides some guidelines on the choice of stopwords list as well, such that the PoS aspect could be taken into account in order to ensure there is sufficient coverage in terms of number of noun and verb entities that trigger nonzero sentence-to-sentence similarity score.
- The findings in Sect. 4 highlight the benefits of the PoS transformation, especially the "all-to-noun" transformation as a tool that can overcome some inherent limitation of the canonical sentence-to-sentence semantic similarity where only noun and verbal entities are handled. Nevertheless, this should not hide the inherent limitations of such an approach which primarily relies on either manually created database or lexical transformation rules, which are not error free, and can lead to an amplification of the meaning shift from the correct sense. The analysis of the computational complexity revealed the importance to account for WordNet server latency and reliability, which is often outside the scope of the user. Therefore, any attempt to use backup and locally stored data instead of server data is of paramount importance in this regard.
- In overall, the development pursued in this paper, despite their novelty and relevance to AI and NLP community, is also subject to several inherent limitations, which can be rooted back to either the selected semantic similarity measures or the generic pipeline employed. First, we restricted our analysis to the three commonly employed WordNet similarity measures that explore the hierarchical structure only, which leaves other similarity measures unexplored, including those proposed as extensions/refinements of path-length and Wu and Palmer measures, see [51] for an overview of structured similarity measures. Second, the exploration of properties at either word level or sentence level is not meant to be exhaustive as we restricted to only few properties that might be of interest to data mining and natural language processing community. Third, the PoS conversion process utilized WordNet's conceptual relations and some other linguistics/grammatical rules. Therefore, such conversation, sometimes, may not be accurate, especially when the underlined word has multiple senses, which opens wide the door of word sense disambiguation problem. Fourth, the extension from word level to sentence level considered only the canonical extension highlighted in expression (21). Nevertheless, it should be noted that this is far to be the unique representation of such extension and several interesting proposals have been reported in the computational linguistic community. One may mention, for instance, approaches that exploit the syntactic information in the sentence linking verb entity to their subject and complement object in a way to maintain up to some extent the structure of the sentence, n-gram models or enforcing some specific sentence ontology representation, see [18] for an overview of such alternative representations.

9 Conclusion and future work

This paper investigates the sentence semantic similarity using the WordNet lexical database where three path-based word similarity measures have been contrasted and investigated. Some theoretical and algebraical properties of the underlined semantic similarities have been examined and scrutinized, providing the user with tremendous help when deciding on the choice of a specific word semantic similarity measure. Similarly, the canonical extension from word-level semantic similarity to sentence-level semantic similarity has been examined with respect to the properties of the underscored extension. It has particularly been highlighted the impact of the PoS tagging on the outcome of the sentence similarity. To handle inherent limitations of word semantic similarity, especially when deficiency in word category is observed in a sentence, a new proposal

using PoS WordNet-based conversion has been put forward. The performance of the proposal in terms of correlation with human judgment has been compared to conventional canonical extension and some state-of-the-art techniques using O'Shea et al. [46] publicly available dataset. The comparative analysis demonstrated the feasibility and superiority of the proposal where using the all-to-noun conversion in the canonical extension expression is found to outperform the conventional canonical extension as well as state-of-the-art STASIS and LSA method. Besides, using the t-test testing, this outperformance is found to be statistically significant at $p < 0.005$ for conventional extension and STASIS. As an immediate future work, it is expected to extend the "All-to-noun" conversion scheme by incorporating ways of handling named entities as this plays special linguistic role, and let the overall sentence similarity be constituted of two parts: one part will be related to standard sentence-level input and another one dedicated to named-entity similarity only. Furthermore, it is also among our plans to emphasize the similarity of sentence's corresponding frame roles using semantic role labelling, which enables us to account to a large extent for the sentence grammar. In the case of lexical resource, we intend to employ ConceptNet database instead of WordNet as part of the future works in our ongoing research.

Appendix

Proof of Proposition 3 The implication $Sim_x(c_i, c_j) = 1 \Leftarrow (c_i = c_j)$ is trivial from the reflexivity property of the three semantic similarity measures. To prove the reverse implication $Sim_x(c_i, c_j) = 1 \Rightarrow c_i = c_j$, one shall proceed for each similarity measure, and noticing that *le* only if $c_i = c_j$.

Using path-length measure, we have:

$$\frac{1}{len(c_i, c_j)} = 1 \Rightarrow len(c_i, c_j) = 1 \Rightarrow c_i = c_j$$

Using normalized Leacock and Chodorow measure, we have from (4),

$\mathrm{Sim}_{\mathrm{lch}}(c_i, c_j) = \log(2{*}\max\_depth)$, So

$$\frac{2{*}\max\_depth}{len(c_i, c_j)} = 2{*}\max\_depth \Rightarrow len(c_i, c_j) = 1 \Rightarrow \mathrm{c_i} = c_j$$

Using Wu and Palmer's measure. Let us assume that c_i, c_j have distinct nodes in the network. Then, let depth(lcs((c_i, c_j)) = l, length(c_i, lcs) = l_1, length(c_j, lcs) = l_2. Therefore, depth(c_i) = l + l_1, depth(c_j) = l + l_2. So,

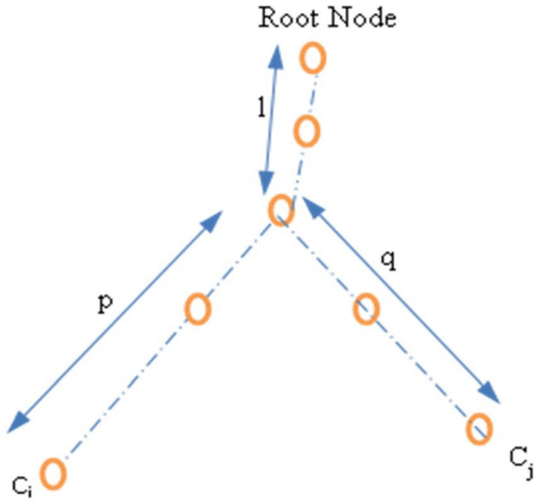

Fig. 3 A taxonomy to two concepts

$$Sim_{wp}(c_i, c_j) = 1 \Rightarrow \frac{2l}{2l + l_1 + l_2} = 1$$
$$\Rightarrow l_1 + l_2 = 0 \Rightarrow (l_1 = 0 \text{ and } l_2 = 0) \Rightarrow \mathrm{c}_i = \mathrm{c}_j$$

Proof of Proposition 5 To prove the statement, let us consider without loss of generality the generic taxonomy of Fig. 3 showing the path between the two synsets c_1 and c_2 as well as their lower common subsumer.

$$\boldsymbol{Sim_{path}(c_i, c_j) = \frac{1}{p+q}, Sim_{wup}(c_i, c_j) = \frac{2l}{2l+p+q}}$$

Since parameters p, q and l are positively valued, it holds that $+q + 2l \geq p + q$, this again entails (since 2 l > l):

$$\frac{1}{p+q+2l} \leq \frac{1}{p+q} \leq \frac{2l}{p+q}$$

Therefore, the inequality $Sim_{path}(c_i, c_j) \leq Sim_{wup}(c_i, c_j)$ trivially holds.

Denoting for simplicity $x = len(c_i, c_j)$, d = max_depth, then $Sim_{wup}(c_i, c_j) \leq Sim^*_{lch}(c_i, c_j)$ is equivalent to $\frac{\log\left(\frac{x}{2d}\right) - \log 2}{\log(2d) - \log 2} \geq \frac{1}{x}$ or, equivalently

$$\frac{-\log x + \log(2d) - \log 2}{\log(2d) + \log 2} - \frac{1}{2} \geq 0 \qquad \text{(A.1)}$$

By derivating (A.1) with respect to x, we have $-\frac{1}{x}\left(\frac{1}{\log(2d) - \log 2} + \frac{1}{x}\right) \geq 0$, which entails

$$x \leq \log(2d) - \log 2 \approx 2.98$$

Proof of Proposition 7 To illustrate the skeleton of the proof, let us consider the generic two examples shown in Fig. 4.

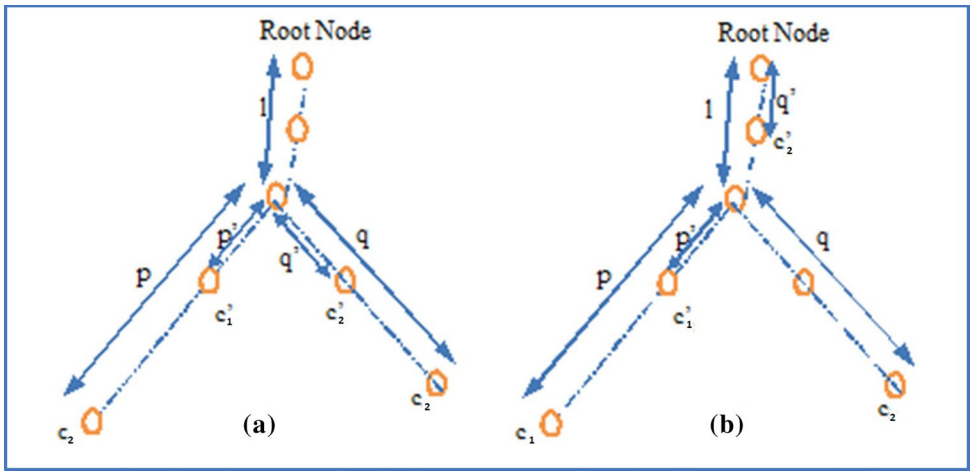

Fig. 4 An example of related synsets

To prove statement in (i), notice that Fig. 4a highlights a typical scenario which c_1, c_2, c_1' and c_2' have the same lower common subsumer. In such case, it holds that

$$len(c_1', c_2') = p' + q' \leq p + q = len(c_1, c_2)$$
$$\Rightarrow Sim_{path}(c_1', c_2') \geq Sim_{path}(c_1, c_2)$$

This also entails $Sim^*_{lch}(c_1', c_2') \geq Sim^*_{lch}(c_1, c_2)$ similarly, we have

$$Sim^*_{wup}(c_1', c_2') = \frac{2l}{p' + q' + 2l} \geq \frac{2l}{p + q + 2l} = Sim_{wup}(c_1, c_2)$$

To prove statement (ii) where synsets are such that c_1' and c_2' are direct hyponyms of c_1 and c_2 without a common lowest common subsumer, one notices that such scenario implicitly entails that either c_1 is lowest common subsume of c_2 or vice versa. For instance, if c_1 is lowest common subsume, the following diagram holds

$$c_2 \rightarrow c_2' \ldots \rightarrow c_1 \rightarrow c_2' \rightarrow ...Root$$

In such case, it holds that

$$len(c_1', c_2') = len(c_1, c_2) \Rightarrow Sim_{path}(c_1', c_2') = Sim_{path}(c_1, c_2)$$

For similar arguments, $Sim^*_{lch}(c_1', c_2') = Sim^*_{lch}(c_1, c_2)$, while

$$Sim_{wp}(c_1', c_2') = \frac{2(l-1)}{len(c_1, c_2) + l - 1 + l - 1}$$
$$= \frac{2(l-1)}{len(c_1, c_2) + 2(l-1)} \geq \frac{2l}{len(c_1, c_2) + 2l}$$
$$= Sim_{wp}(c_1, c_2)$$

So, in both cases it holds that $Sim_*(c_1, c_2) \geq Sim_*(c_1', c_2')$.

To prove (iii), it is enough to see Fig. 4b, where $len(c_1', c_2') = p' + l - q'$ while $len(c_1, c_2) = p + q$. Since l is fully independent of q, $len(c_1', c_2')$ can be greater, equal or smaller than $len(c_1, c_2)$ so that no specific ordering can be established. Same reasoning applies when calculating the depth of the synsets, which renders $Sim_*(c_1, c_2)$ and $Sim_*(c_1', c_2')$ not comparable.

Proof of Proposition 8

(i) From the assumption that c_i is a direct hyponym of c_j, it follows c_i is also the least common subsumer of the two synsets. So, if depth(c_i)=l, then depth (c_j)=l+1. Therefore, $Sim_{wp}(c_i, c_j) = \frac{2l}{2l+1}$. Noticing that the above expression is non-decreasing in l, and for distinct synsets, the minimum value of l is 2, which, after substituting in the above expression, yields $Sim_{wp}(c_i, c_j) = 0.8$. The result follows straightforwardly that if c_i is a direct hyponym ofc_j, then len(c_i, c_j)=2, so after substituting in (3) and (4), the result (ii) and (iii) of Proposition 8 are trivial.

Proof of Proposition 9 The hyponymy relation can be represented by the diagram below $c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow \ldots$ $c_{n-1} \rightarrow c_n \rightarrow \ldots$ *RootNode*.

Given that$len(c_1, c_2) = 2 \leq len(c_1, c_3) = 3 \leq \ldots \leq .len(c_1, c_k) = k$ *for* $k4, n$. This indicates that (19) trivially holds for path and Leacock and Chodorow similarity. For Wu and Palmer similarity, assume a length l from c_n till *RootNode*, then it holds.

$$Sim(c_1, c_2) = \frac{2(l+n-1)}{l+n+l+n-1} = \frac{2(l+n-1)}{2(l+n-1)-1} = \frac{1}{1+\frac{1}{2(l+n-1)}}$$

While

$$Sim(c_1, c_k) = \frac{2(l+n-k-1)}{l+n+l+n-k-1} = \frac{2(l+n-k-1)}{2(l+n-k-1)+k+1} = \frac{1}{1+\frac{k+1}{2(l+n-k-1)}},\ \text{k = 3,n}$$

Noticing that $\frac{1}{2(l+n-1)} < \frac{k+1}{2(l+n-k-1)}$ since this is equivalent to$[2l-2(k+1)l]+[2n-2(k+1)n]+[-2-2(k+1)]<0$, which trivially holds since each expression under square bracket on the left hand side of the last inequality is always negatively valued for k greater or equal than 3. This yields $Sim_*(c_1, c_2) \geq Sim_*(c_1, c_k)$ for k = to n, which, by a simple induction reasoning, yields inequality (20).

Proof of Proposition 11 Applying expression (21) with respect to the above-defined S_1 and S_2 leads to, after noticing, that due to semantic equivalence assumption $Sim_*(N_{1i}, N_{2j}) = Sim_*(N_{11}, N_{21})$, $j=1$ to s, $i=1$ to p, and $Sim_*(V_{1i}, V_{2j}) = Sim_*(V_{11}, V_{21})$, j = 1 to l, i = 1 to q, we have

$$Sim_g^*(S_1, S_2) = \frac{1}{2}\left[\frac{pSim_*(N_{11}, N_{21}) + qSim_*(V_{11}, V_{21})}{p+q} + \frac{sSim_*(N_{21}, N_{11}) + lSim_*(V_{21}, V_{11})}{s+l}\right]$$

$$= \frac{1}{2}\left[Sim_*(N_{11}, N_{21})\left(\frac{p}{p+q}+\frac{s}{s+l}\right) + Sim_*(V_{11}, V_{21})\left(\frac{q}{p+q}+\frac{l}{s+l}\right)\right]$$

Noticing that $Sim_g(S_1', S_2')$ would have the same form as (27), reconciling the last expression with that of (27) yields

$$\frac{p}{p+q}+\frac{s}{s+l} = 1 \text{ and } \frac{q}{p+q}+\frac{l}{s+l} = 1$$

After some manipulations, the last expressions are equivalent to $ps = ql$, which is again equivalent to $\frac{p}{q} = \frac{l}{s}$.

Proof of Proposition 13 From the definition of S_A and S_B, it holds

$$Sim_g^*(S_A, S_B) = \frac{\sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}') + Sim_*(N_{k1}, N_{p+1,2}) + \sum_{i=1}^{q} Sim_*(V_{i1}, V_{i2}')}{p+q}$$

Similarly,

$$Sim_g^*(S_A, S_B') = \frac{\sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}') + Sim_*(N_{k1}, N_{p+1,2}) + \sum_{i=1}^{q} Sim_*(V_{i1}, V_{i2}')}{2(p+q)} + \frac{\sum_{i=1}^{p} Sim_*(N_{i1}, N_{i2}') + Sim_*(N_{k1}, N_{p+1,2}) + \sum_{i=1}^{q} Sim_*(V_{i1}, V_{i2}')}{2(p+q+1)}$$

In the first fraction of $Sim_g(S_A, S_B')$, it assumes that $N_{p+1,2}$ does not influence the semantic similarity of other pairs. In the opposite case, the overall still is valid as will be demonstrated later on. By rewriting.

$$\sum_{i=1}^{p} Sim_*(N_{i1}, N_{i2}') = \sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}') + Sim_*(N_{k1}, N_{k2}')$$

And after some manipulations, subtracting $Sim_g(S_A, S_B)$ from $Sim_g(S_A, S_B')$, we have

$$Sim_g^*(S_A, S_B') - Sim_g(S_A, S_B) \propto$$

$$(2p+2q+l)\sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}') + (2p+2q+1)\sum_{i=1}^{q} Sim_*(V_{i1}, V_{i2})$$

$$+(2p+2q+1)Sim_*(N_{k1}, N_{p+1,2}) + (p+q)Sim_*(N_{k1}, N_{k2}')$$

$$-2(p+q+1)\sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}')$$

$$-2(p+q+1)\sum_{i=1, i\neq k}^{q} Sim_*(V_{i1}, V_{i2})$$

$$-(2p+2q+1)Sim_*(N_{k1}, N_{k2}')$$

This can be simplified as $Sim_g(S_A, S_B') - Sim_g(S_A, S_B)$

$$\propto (2p+2q+1)Sim_*(N_{k1}, N_{p+1,2}) - \sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}') - \sum_{i=1}^{q} Sim_*(V_{i1}, V_{i2}') - (p+q)Sim_*(N_{k1}, N_{k2}')$$

The last expression can be rewritten as

$$\left[(p-1)Sim_*(N_{k1}, N_{p+1,2}) - \sum_{i=1, i\neq k}^{p} Sim_*(N_{i1}, N_{i2}')\right]$$

$$+\left[qSim_*(N_{k1},N_{p+1,2})-\sum_{i=1,}^{q}Sim_*(V_{i1},V'_{i2})\right]$$

$$+(p+q+2)\left[Sim_*(N_{k1},N'_{k2})\right]-Sim_*(N_{k1},N'_{k2})$$

It is then easy to see that each expression under square bracket in above is positively valued as $Sim_*(N_{k1},N_{p+\_1,2})$ is supremum over semantic similarity of any chosen pairs belonging to either noun or verb part of speech. This makes

$$Sim_g^*(S_A,S'_B)-Sim_g^*(S_A,S_B)>0$$

In case where $N_{p+1,2}$ does influence the semantic similarity of other pairs, say, there is a noun N_{m1} in sentence S_A such that $Sim_*(N_{m1},N'_{m2})\le Sim_*\left(N_{m1},N'_{p+1,2}\right), Sim_g(S_A,S'_B)$ is changed as ($Sim_g^*(S_A,S_B)$ remains unchanged):

$$Sim_g^*(S_A,S'_B)=\frac{\sum_{i=1,i\neq k}^{p}Sim_*(N_{i1},N'_{i2})+Sim_*(N_{k1},N_{p+1,2})+\sum_{i=1}^{q}Sim_*(V_{i1},V'_{i2})}{2(p+q)}$$
$$+\frac{\sum_{i=1}^{p}Sim_*(N_{i1},N'_{i2})+Sim_*(N_{k1},N_{p+1,2})+\sum_{i=1}^{q}Sim_*(V_{i1},V'_{i2})}{2(p+q+1)}$$

Using similar reasoning as above, the quantity $Sim_g^*(S_A,S'_B)-Sim_g^*(S_A,S_B)$ is proportional to

$$\left[(p-2)Sim_*(N_{k1},N_{p+1,2})-\sum_{i=1,i\neq k,i\neq m}^{p}Sim_*(N_{i1},N'_{i2})\right]$$
$$+\left[(qSim_*(N_{k1},N_{p+1,2})-\sum_{i=1}^{q}Sim_*(V_{i1},V'_{i2})\right]$$
$$+(p+q+1)\left[Sim_*(N_{m1},N_{p+1,2})-Sim_*(N_{m1},N'_{m2})\right]$$
$$+(p+q+2)\left[Sim_*(N_{k1},N_{p+1,2})-Sim_*(N_{k1},N'_{k2})\right]$$
$$+\left[Sim_*(N_{k1},N_{p+1,2})-Sim_*(N_{m1},N'_{m2})\right],$$

where each entity under bracket is also positively valued.

The same reasoning applies if there is more than only element N_{i1} in sentence S_A, for which the inequality $Sim_*(N_{i1},N'_{i2})\le Sim_*(N_{i1},N_{p+1,2})$ is held. The details is omitted for its similarity with previous one. This completes the proof.

Acknowledgements This work is partly supported by EU YoungRes project (#823701), which is gratefully acknowledged.

Funding Open access funding provided by University of Oulu including Oulu University Hospital.

References

1. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
2. Fellbaum, C., *WordNet: An Electronic Lexical Database*, MIT Press, 1998
3. Banches, R.E.: Text Mining With Matlab. Springer, NY (2013)
4. Feldman R. and Sanger J., *The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press NY, 2007
5. Ali M., Ghosh M., and Al-Mamun A., "Multi-document text summarization: Simwithfirst based features and sentence co-selection based evaluation," in: *International Conference on Future Computer and Communication (ICFCC)*, 2009, pp. 93–96
6. Achananuparp P., Hu X., Zhou X., and Zhang X., "Utilizing sentence similarity and question type similarity to response to similar questions in knowledge-sharing community," in: *Proceedings of QAWeb* Workshop, 2008
7. Cremmins, E.T.: The Art of Abstracting. Information Resources Press, Arlington, VA, second edition (1996)
8. Allan J., Bolivar A., and C. Wade, Retrieval and novelty detection at the sentence level. In: *Proceedings of SIGIR'03*, 2003, pp. 314–321
9. Balasubramanian, N., Allan, J., and Croft, W. B., A comparison of sentence retrieval techniques. In: *Proceedings of SIGIR '07*, Amsterdam, the Netherlands, 2007, 813–814
10. Osman, A., Salim, N., Binwahlan, M., Twaha, S., Kumar, Y., and Abuobieda, A., "Plagiarism detection scheme based on semantic role labeling,"in: *International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, pp. 30–33
11. Hoad, T., Zobel, J.: Methods for identifying versioned and plagiarized documents. J. Am. Soc. Inf. Sci. Technol. **54**(3), 203–215 (2003)

12. Jun-Peng, B., Jun-Yi, S., Xiao-Dong, L., Qin-Bao, S.: A survey on natural language text copy detection. J. Softw. **14**(10), 1753–1760 (2003)
13. Haque. R., Naskar, S., Way, A., Costa-jussa, M., and Banchs, R., "Sentence similarity-based source context modelling in pbsmt," in: *International Conference on Asian Language Processing (IALP)*, 2010, pp. 257–260
14. Ko, Y., Park, J., Seo, J.: Improving text categorization using the importance of sentences. Inf. Process. Manage. **40**(1), 65–79 (2004)
15. Madhavan, J., Bernstein, P., Doan, A. , Helevy, A., "Corpus-based schema matching". In: *Procedings of the International Conference on Data Engineering*, 2005
16. Chiang, J.H., Yu, H.C.: Literature extraction of protein functions using sentence pattern mining. IEEE Trans. Knowledge and Data Eng. **17**(8), 1088–1098 (2005)
17. Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys.-Doklady **10**, 707–710 (1966)
18. Chandrasekaran, D. and Mago, V., "Evolution of Semantic Similarity", *ACM Comput. Surveys*, 54(2), 2021, 41:1–41:37
19. Chersoni, E., Santus, E., Pannitto, L., Lenci, A., Blache, P., Huang, C.: A structured distributional model of sentence meaning and processing. Nat. Lang. Eng. **25**(4), 483–502 (2019)
20. Navigli, R.: Word sense disambiguation: a survey. ACM Computational Survey **41**(2), 1–69 (2009)
21. Quan, Z., Wang, Z., Le, Y., Yao, B., Li, K., Yin, J.: An efficient framework for sentence similarity modeling. IEEE/ACM Trans. Audio, Speech, Language Process. **27**(4), 853–865 (2019)
22. Liu, H., Singh, P.: ConceptNet a practical commonsense reasoning tool-kit. BT Technol. J. **22**(4), 211–226 (2004)
23. Mihalcea, R., Corley, C., and Strapparava, C., "Corpus-based and knowledge-based measures of text semantic similarity," in: *Proceedings of AAAI*, Vol. 6, 2006, pp. 775–780
24. Bawakid, A. and Oussalah, M., "A semantic-based text classification system," in: *Cybernetic Intelligent Systems (CIS), IEEE 9th International Conference*, 2010, pp. 1–6
25. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowledge Data Eng. **18**(8), 1138–1150 (2006)
26. Li, Y.H., Bandar, Z., McLean, D.: An approach for measuring semantic similarity using multiple information sources. IEEE Trans. Knowledge Data Eng. **15**(4), 871–882 (2003)
27. Islam, A. and Inkpen, D., "Semantic text similarity using corpus based word similarity and string similarity", *ACM Transactions on Knowledge Discovery from Data*, 2(2), 2008, 10:1–10:25
28. Ozates¸ S. A-B., Ozgu,r A., Radev, D. R., "Sentence Similarity based on Dependency Tree Kernels for Multi-document Summarization", In: *Proceedings of LREC 2016*, pp. 2833–2838
29. Wu, Z. and Palmer, M., "Verbs semantics and lexical selection", in: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138
30. Leacock, C. and Chodorow, M., "Combining local context and WordNet similarity for word sense identification," In: *WordNet: An Electronic Lexical Database*, The MIT Press, 1998, pp. 265–283
31. Resnik, P., "Using information content to evaluate semantic similarity," In: *Proceedings of the 14th Int Joint Conf. on Artificial Intelligence*, 1995, pp. 448–453
32. Lin, D., "An information-theoretic definition of similarity," In: *Proceedings Of the International Conference on Machine Learning*, 1998, pp. 296–304
33. Jiang, J.J. and Conrath, D.W., "Semantic similarity based on corpus statistics and lexical taxonomy," In: *Proceedings of the International Conference on Research in Computing Linguistics*, 1997, pp. 19–33
34. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G. M. and Milios, E. E., "Semantic similarity methods in WordNet and their application to information retrieval on the web", *Proceedings of the 7th annual ACM international workshop on Web information and data management,* Bremen, Germany, 2005, pp.10–16
35. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance. Comput. Linguist. **32**(1), 13–47 (2006)
36. Meng, L., Huang, R., Gu, J., "A review of semantic similarity measures in WordNet," *International Journal of Hybrid Information Technology*, 6(1), 2013
37. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**(1), 17–30 (1989)
38. Mohamed, M., Oussalah, M.: SRL-ESA-TextSum: a text summarization approach based on semantic role labeling and explicit semantic analysis. Inf. Process. Manage. (2019). https://doi.org/10.1016/j.ipm.2019.04.003
39. Mohamed, M., Oussalah, M.: A hybrid approach for paraphrase identification based on knowledge-enriched semantic heuristics. Lang. Resour. Eval. (2019). https://doi.org/10.1007/s10579-019-09466-4
40. Malik, R., Subramaniam, L. V., and Kaushik, S., "Automatically selecting answer templates to respond to customer emails." in *Proceedings of the IJCAI*, vol. 7, 2007, pp. 1659–1664
41. Mohamed, M. and Oussalah, M., "A comparison study of conversion aided methods for WorldNet sentence textual similarity", in: Proceedings of Workshop in Information Discovery in Text, *25th International Conference on Computational Linguistics*, Ireland, 2014
42. Borwein, J. and Borwein, P., *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*. John Wiley, 1987
43. Habash, N. and Dorr B., "A categorial variation database for English," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*,Volume 1. Association for Computational Linguistics, 2003, pp. 17–23
44. Miller, G.A., Fellbaum, C.: Morphosemantic links in WordNet. Traitement automatique des Langues **44**(2), 69–80 (2003)
45. Morsphosemantic-links, available at http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls
46. O'Shea, J., Bandar, Z., Crockett K., McLean D., "A comparative study of two short text semantic similarity measures," in Agent and Multi-Agent Systems: Technologies and Applications, ed: Springer, 2008, pp. 172–181
47. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**, 391 (1990)
48. Pedersen, T., Patwardhan, S., and Michelizzi, J., "Wordnet:: Similarity: measuring the relatedness of concepts," in: *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 38–41
49. Pedersen, T., WordNet Similarity Measure module, available at http://sourceforge.net/projects/wn-similarity/
50. Illinois Part of Speech Tagger. Available in http//cogcomp.cs.illinois.edu/ page/software view/POS)
51. Sebastien, H., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis. Synthesis Lectures Human Language Technol. **8**(1), 1–254 (2015)