The Knowledge Engineering Review, page 1 of 30. © Cambridge University Press, 2017
doi:10.1017/S0269888917000029

The state of the art in semantic relatedness: a framework for comparison

YUE FENG[1], EBRAHIM BAGHERI[1], FAEZEH ENSAN[2] and JELENA JOVANOVIC[3]

[1]*Laboratory for Systems, Software and Semantics (LS^3), Ryerson University, Toronto, M5B 2K3 ON, Canada;*
e-mail: bagheri@ryerson.ca;
[2]*Department of Computer Engineering, Ferdowsi University of Mashhad, Azadi Square, Mashhad, Razavi Khorasan*
e-mail: ensan@um.ac.ir;
[3]*Department of Software Engineering, School of Business Administration, University of Belgrade, Jove Ilica 154, 11000 Belgrade, Serbia;*
e-mail: jeljov@fon.rs

Abstract

Semantic relatedness (SR) is a form of measurement that quantitatively identifies the relationship between two words or concepts based on the similarity or closeness of their meaning. In the recent years, there have been noteworthy efforts to compute SR between pairs of words or concepts by exploiting various knowledge resources such as linguistically structured (e.g. WordNet) and collaboratively developed knowledge bases (e.g. Wikipedia), among others. The existing approaches rely on different methods for utilizing these knowledge resources, for instance, methods that depend on the path between two words, or a vector representation of the word descriptions. The purpose of this paper is to review and present the state of the art in SR research through a hierarchical framework. The dimensions of the proposed framework cover three main aspects of SR approaches including the resources they rely on, the computational methods applied on the resources for developing a relatedness metric, and the evaluation models that are used for measuring their effectiveness. We have selected 14 representative SR approaches to be analyzed using our framework. We compare and critically review each of them through the dimensions of our framework, thus, identifying strengths and weaknesses of each approach. In addition, we provide guidelines for researchers and practitioners on how to select the most relevant SR method for their purpose. Finally, based on the comparative analysis of the reviewed relatedness measures, we identify existing challenges and potentially valuable future research directions in this domain.

1 Introduction

Humans can often effortlessly decide about the similarity or relatedness of two words[1]. This can be explained, in part, by the experience that humans have in using and encountering related words in similar contexts. For instance, as human beings, we know *rain* and *umbrella* are highly related, while there is a little, if any, connection between *rain* and *textbook*. While this is trivial for humans, it is often not as simple to translate this judgment process for machines without the careful formulation of background and contextual knowledge surrounding each word and its relationships. Formally speaking, semantic relatedness (SR) is defined as a form of semantic or functional associations between two words rather than just lexical relations such as synonymy and hyponymy (Budan & Graeme, 2006). The objective of SR methods is to closely model such associations.

SR is widely used in many practical applications, particularly in natural language processing (NLP) including semantic information retrieval, keyword extraction and document summarization, where it is

[1] While acknowledging the differences, we use the terms ‘words, concepts, terms and entities’, interchangeably in this paper.

used to quantify the relations between words or between words and documents (Leong & Mihalcea, 2011). Information retrieval techniques have particular interest in SR measures as their incorporation in the retrieval process allows for the identification of meaningfully related but lexically dissimilar content (Budanitsky & Hirst, 2006). Other more specialized domains such as biomedical informatics and geoinformatics have also benefited from SR techniques to identify relationships between bioentities (Pedersen *et al*., 2007) and geographic concepts (Hecht *et al*., 2012), respectively.

The development and formalization of SR methods is a formidable task that requires solutions for various challenges. In this paper, we are primarily concerned with two main challenges in this area: (1) challenges related to the underlying knowledge resources that can provide insight into SR of words, and (2) challenges related to the formalization of the relatedness measures. In order to understand the scope of these two challenges and to identify the current state of the art, we extensively review work in the area of SR, specifically attempting to cover the main models and techniques that have been proposed to address each of the two challenges.

To this end, we propose a taxonomic framework for comparing some of the more widely known work in this domain with specific focus on the above two aspects. The framework is presented by considering the basic features of SR methods including: (i) the knowledge resources that an SR method adopts; (ii) the computational methods that an SR method is based on; and (iii) the evaluation method that is used to assess the suitability of an SR method, including the used data sets and evaluation metrics. The framework dimensions and its sub-dimensions are used as a basis for comparing the strengths and weaknesses of the wider known work in the domain; consequently, providing a guideline for researchers and practitioners to choose appropriate features when constructing or selecting SR methods according to their needs.

The rest of this paper is organized as follows: Section 2 clearly outlines the criteria used for selecting the methods studied in this paper and describes each method in detail. Section 3 presents the proposed framework, and its dimensions and sub-dimensions. Section 4 compares the selected methods and discusses the strengths and weakness of each method in the context of the proposed framework. Section 5 provides a meta-analysis of the findings in this paper and identifies areas for future research. Section 6 concludes the paper.

2 Semantic relatedness methods

Our research objective is to develop a framework that allows us to compare some of the well-known methods in the SR literature. We adopted an iterative approach towards the design of this framework. We initially base our work on the three main dimensions that have already been highlighted in the literature (Agirre *et al*., 2009; Chen *et al*., 2009), namely knowledge resources, computational methods and evaluation approaches. We then identify several important work in the literature that would be considered seminal or novel work in the domain of SR. Our criteria for selecting these methods are as follows:

- *Selecting methods with a substantial impact on the literature*: Our objective has been to select and review methods that have had a notable impact on the research community. For this purpose, one of the criteria for choosing a study has been its citation count obtained through Google Scholar. We postulate that the higher the citation count for a publication, the better the proposed method has been received and recognized by the community.
- *Selecting methods with original proposals*: Our goal has been to include work that was the first to propose an idea with regards to using a knowledge resource or a computation method. The selection included studies that were original work in proposing the idea and not adoptions of earlier ideas. To decide on originality of two similar pieces of work, the work published earlier and cited by other work in an earlier chronological order was chosen as the original one.

For example, we chose Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2007) since it is the pioneering work in exploring Wikipedia's articles and concepts as the underlying knowledge resource. Another example is the work by Sahami and Heilman (2006) who was one of the first to propose the use of Web search engine results for developing a similarity kernel function. Table 1 shows a summary of the selected methods ordered by their citation counts along with their references, and a brief introduction of

https://doi.org/10.1017/S0269888917000029

Table 1 Summary of selected methods

Method name	References	Citation count	Brief description
Resnik	Resnik (1995)	2675	Considers the information content value of two words based on subsumption relations in a taxonomy
Jiang and Conrath	Jiang and Conrath (1997)	2331	Considers the information content value of two words' subsumption relations as well as the information content value of two words in a taxonomy
Lesk	Lesk (1986)	1506	Computes the amount of word overlap between the glosses of each word pair
ESA	Gabrilovich and Markovitch (2007)	1189	Generates a concept vector to represent each word by exploring related Wikipedia articles
Cilibrasi and Vitányi	Cilibrasi and Vitanyi (2007)	1138	Considers the number of pages returned by Google in which the two words co-occur
WikiRelate!	Strube and Ponzetto (2006)	623	Calculates the length of a path between two nodes in the graph constructed by Wikipedia's articles and category tree
Sahami and Heilman	Sahami and Heilman (2006)	518	Mines additional information from public Web pages to enhance the representation of a word
Patwardhan and Pedersen	Patwardhan and Pedersen (2006)	275	Constructs a second-order gloss vector for each word from Wordnet
Hughes and Ramage	Hughes and Ramage (2007)	133	Applies random walk on the graph constructed by exploring the relationship structure of Wordnet
TSA	Radinsky *et al*. (2011)	101	Creates a time series concept vector to represent each word by exploring related articles' history in Wikipedia
WLM	Milne (2007)	99	Constructs vectors for each word by using the links in Wikipedia articles
Zesch *et al*.	Zesch *et al*. (2008)	93	Represents a word using content gathered from the collaboratively constructed dictionary *Wiktionary*
Gur	Gurevych (2005)	58	Constructs pseudo-glosses for each word by concatenating concepts in close relationship with the word
REWOrD[1]	Pirró (2012)	4	Makes use of predicates from semantic Web resources to represent a word

ESA = Explicit Semantic Analysis; TSA = Temporal Semantic Analysis.

[1] This paper has few citations, but it is the only method which uses predicates in semantic resources at the time of writing this paper.

each method. In order to avoid papers with high citation count due to self-citations or semi-self-citations, the citations of the selected papers were manually reviewed.

Once the methods were selected based on the above criteria, the approaches introduced in each of these 14 methods were studied with respect to the three main dimensions of our taxonomy. For instance, the fact that Resnik (1995) uses WordNet as the knowledge resource, whereas ESA (Gabrilovich & Markovitch, 2007) employs Wikipedia. Based on the extracted information from each of the 14 methods, we generalized the findings to build a hierarchical representation of the types of work done in the literature, as shown in Figure 1. Therefore, the development of the taxonomy has been based on evidence derived from the relevant literature and serves as a platform for comparing some of the well-known SR methods. As will be discussed in the remainder of the paper and shown in Tables 4, 5 and 7, the taxonomy shows the diversity of the types of knowledge resources, computational methods and evaluation approaches that have been presented in the literature for measuring SR.

As shown in Table 1, the 14 selected methods utilize a wide range of knowledge resources that have been proposed for SR calculation such as Wikipedia, Web search engines, ontologies and Wiktionary, to name a few. Furthermore, they cover the state of the art computational methods that are based on word co-occurrences, vector space representation, random walk, path between words or temporal relation between words. In the remaining part of this section, we give an overview of each selected method.

Resnik (1995) hypothesizes that SR between two words is a measure of the amount of information they share. For this purpose, and in order to identify shared information, the method proposed in Resnik (1995) identifies the lowest common subsumer of the two words within an IS-A hierarchy. The information content value of the subsumer is regarded as an indicator of SR.

Jiang and Conrath (1997) employ the information content value of words as well as the information content value of the two words' lowest common subsumer in a lexical taxonomy structure to compute SR. The information content value of two words' lowest common subsumer describes the amount of information these two words share, whereas the information content value of a word indicates how informative that specific word is. Here, SR is defined based on the information content of the lowest common subsumer in the context of the information content of each individual word.

Lesk (1986) structures his work on the short pieces of text (glosses) defining each word in WordNet. Specifically, SR is computed by counting the number of word overlaps in the glosses of the two words, where higher overlap means higher relatedness between two words.

ESA, proposed by Gabrilovich and Markovitch (2007), uses Wikipedia as its underlying knowledge resource. The motivation behind ESA is that Wikipedia contains numerous articles, each one focusing on a single concept; hence, Wikipedia can be viewed as a collection of concepts, each with an article explicitly defined by humans. In their approach, a semantic interpreter is built to map a word into a vector of Wikipedia concepts coupled with weights, where the weights are term frequency-inverse document frequence (TF–IDF) values of the input word in the underlying articles. In this context, SR is measured based on the cosine similarity of the two words' vectors.

Cilibrasi and Vitányi (2007) have proposed a method that relies on the information retrieved from a Web search engine. The motivation behind their work is that similar words when used as search queries will result in similar Web page results. Therefore, the count of the number of shared Web pages returned by a Web search engine for three different search queries, namely w_1, w_2, w_1 and w_2, is used to formalize the *normalized Google distance* (NGD). SR is defined as the inverse of NGD.

WikiRelate! (Strube & Ponzetto, 2006) takes advantage of Wikipedia articles and category tree to compute SR. In their work, the authors apply to Wikipedia the measures that were originally designed for WordNet. Articles are retrieved from Wikipedia by querying word pairs. Wikipedia's disambiguation pages obtained for each word are used for disambiguation of the words. The categories related to the retrieved articles are used to compute SR by for instance, considering the length of the shortest path or the length of the path that maximizes information content.

Sahami and Heilman (2006) have introduced a new approach for computing SR aimed at overcoming the poor performance of traditional document similarity methods when applied on short text snippets (Sahami & Heilman, 2006). Their method, similar to the work in Cilibrasi and Vitanyi (2007), benefit from Web search results. In particular, they leverage Web search results for enhancing short snippets. Top ranked words, based on the TF–IDF measure from the search results, are used to build a vector for each input word. The vector is then used to compute the degree of SR between the two words.

Patwardhan and Pedersen (2006) used the co-occurrence information as well as the definitions of words in WordNet to build gloss vectors corresponding to each word. The gloss vector is created in two steps: (1) the first-order vector consisting of co-occurrences between the target word and other words among all the glosses in WordNet is formed; (2) additionally word co-occurrence information are calculated by concatenating the glosses of words that are related to each other within WordNet. Cosine similarity is applied to the gloss vectors to measure the relatedness between two words.

Hughes and Ramage (2007) present an application of Markov chain theory to measure SR based on a graph extracted from WordNet. The graph is constructed such that the nodes are entries in WordNet and the edges are relational links between words. The authors adopted three types of nodes including *Synset* nodes, *TokenPOS* nodes and *Token* nodes, whereas the relationship types are hypernym/hyponym, instance/instance of, antonym, entails/entailed by, adjective satellite and causes/caused by. SR is calculated by assuming a particle that starts from a specific word, and then roams through the constructed graph. The particle tends to explore the neighborhood related to the target word, hence resulting in a stationary distribution. SR is the similarity between two stationary distributions obtained for the two words.

Temporal Semantic Analysis (TSA) method considers temporal information of resources. It was proposed by Radinsky *et al.* (2011) who hypothesized that by studying the similarity of word usage patterns over time, a great deal of relatedness information can be discovered to enhance the SR results. In their method, each word is represented as a weighted vector of concept time series derived from a historical archive such as *NY Times* archive. Then SR of a pair of words is computed by finding the similarity between two times series representing two words.

WLM: In order to reduce the computation expenses of the ESA approach, Milne (2007) developed a more efficient method by incorporating links found within Wikipedia articles corresponding to the words being compared. The method assumes that the more links two articles share, the more related they are. Thus, a word is represented as vector of links. The links are weighted based on a simple but intuitive idea: articles that receive many incoming links can be considered general articles providing less specific information. Semantic similarity of two words is then the cosine similarity between the weighted vectors representing two words.

Zesch *et al.* (2008) have systematically studied the applicability of Wiktionary as a lexical resource for computing SR. They explored the features of Wiktionary including its relation types, languages, size, instance structure and instance incompleteness in order to propose two SR measures namely a path-based approach and a vector-based approach, explained in detail later in the paper.

Gur: The work by Gurevych (2005) relies on the structure of GermaNet, a conceptual network of relations between German words. Since GermaNet does not include word glosses for word definitions, Gurevych generated artificial conceptual glosses (pseudo-glosses) to describe each word. The pseudo-glosses are constructed by concatenating words that are in close relation to the target words through relations such as synonymy, hypernymy and meronymy, to name a few. SR between two words is then defined as the amount of word overlap between their pseudo-glosses.

REWOrD exploits SPARQL queries to access RDF data from DBPedia and evaluates the relatedness of two words based on the informativeness of the path between the two words (Pirró, 2012). The first step in applying REWOrD is to find DBpedia triples that are relevant to each of the words. The authors then introduce the notion of *informativeness*, which is calculated based on predicate frequency and inverse triple frequency. This is then used to build a vector for each word. The cosine similarity between the vectors for the two words is regarded as their degree of relatedness.

3 Dimensions of the framework

As discussed in the previous section, there are several SR approaches and systems in the literature that differ from each other in the way they approach and define relatedness or the resources they use. In this section, we describe various aspects of SR techniques based on the proposed classification framework, which consists of three main dimensions and several sub-dimensions. The framework dimensions (Figure 1) are as follows:

1. *Knowledge resources*, including:
   a. Linguistically constructed resources (Relations, Synsets in WordNet and GermaNet).
   b. Collaboratively constructed resources (Articles, Article links, Categories, Disambiguation pages in Wikipedia, Information and Relations in English and German Wiktionary).
   c. Web-based resources (Web Search Engines such as Google, Yahoo, Bing and the Semantic Web, i.e. the Linked Open Data cloud).

https://doi.org/10.1017/S0269888917000029

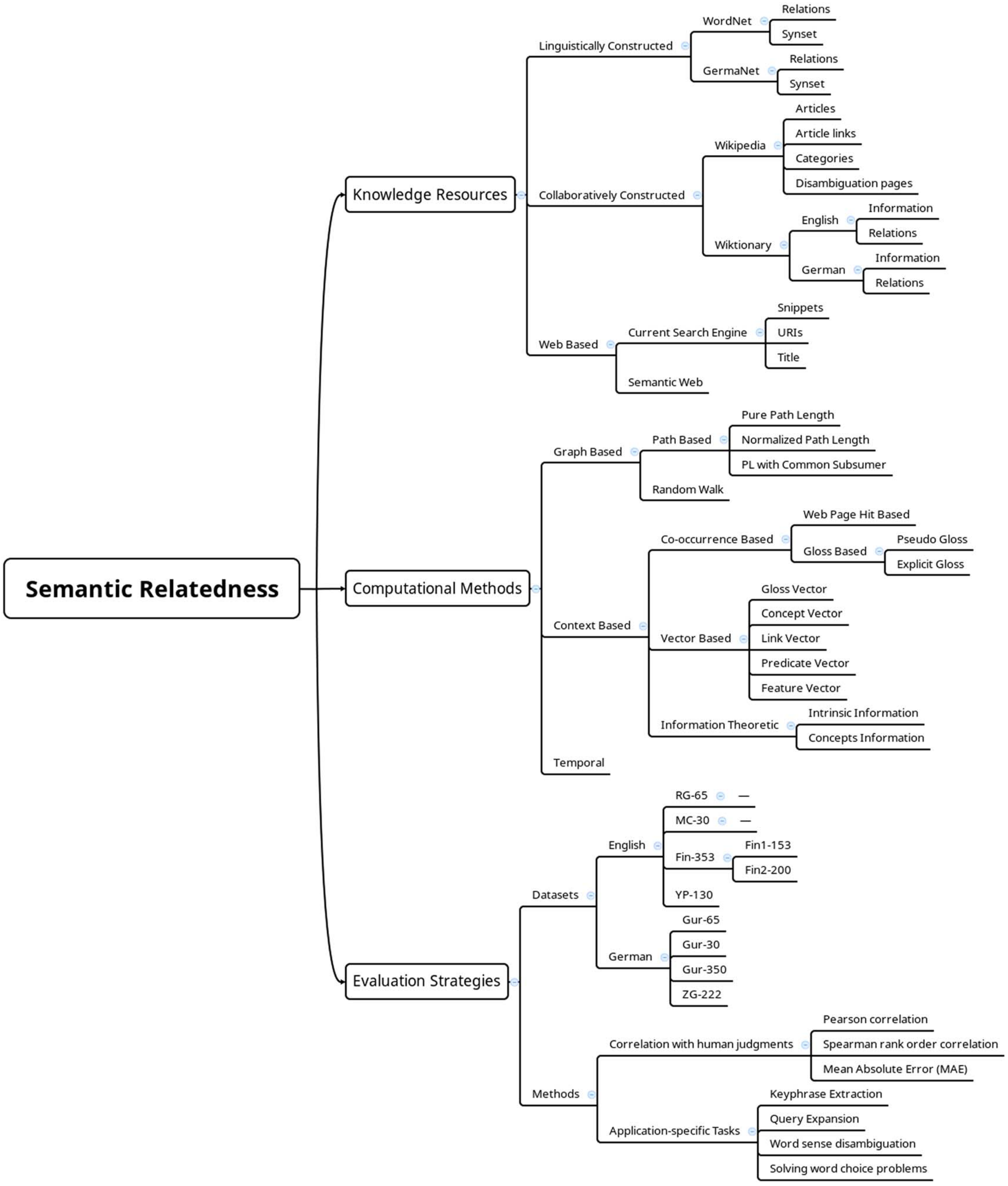

Figure 1 The proposed framework dimensions

2. *Computational methods*, including:
    a. Graph-based methods (Path-based such as Pure Path Length, Graph Length, Common Subsumer and Random Walk).
    b. Context-based methods (Co-occurrence-based such as Web Page Hit; Implicit Gloss or Explicit Gloss-based; Vector-based including Gloss Vector, Concept Vector, Links Vector, Predicates Vector and Feature Vector; Information Content-based such as Concepts Information and Intrinsic Information).
    c. Temporal methods.
3. *Evaluation strategies*, including:
    a. Data sets (English Data sets such as RG-65, MC-30, Fin-353 (Fin1-153,Fin2-200) and YP-130. German Data set like Gur-65, Gur-30, Gur-350 and ZG-222).

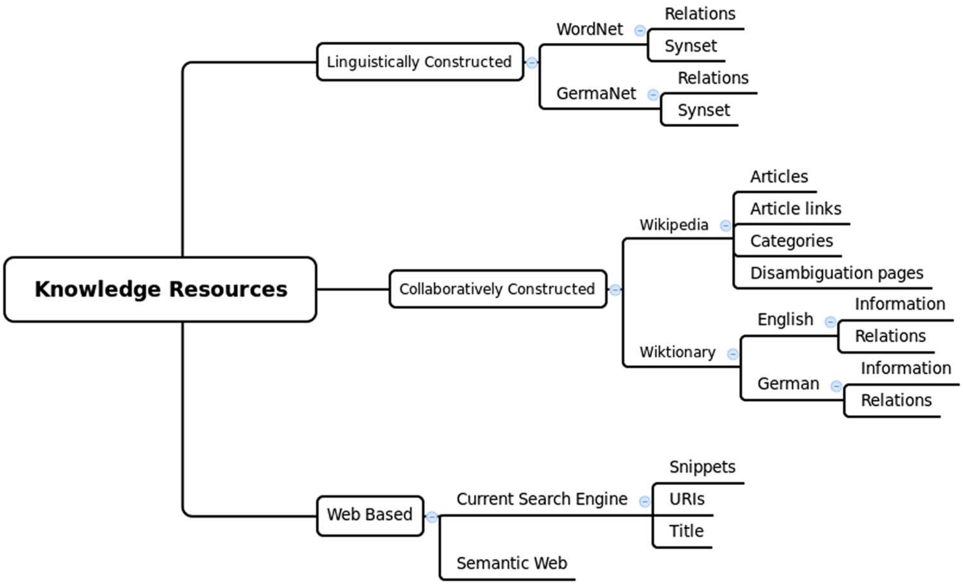

Figure 2 Knowledge resources taxonomy

b. Methods (Correlation with human judgments through Pearson's Correlation or Spearman's Rank Order Correlation. Application specific such as Keyphrase Extraction, Semantic Information Retrieval, Word Sense Disambiguation and Solving word choice problems).

The following subsections describe the framework more deeply covering the three top-level dimensions (Resources, Methods and Evaluation Strategies) and each of their sub-dimensions.

3.1 Knowledge resources

In the context of SR techniques, the term knowledge resource refers to the type and source of information that is used for determining the degree of relatedness between two words. We cover three main types of knowledge resources, namely (i) linguistically constructed resources such as Wordnet, (ii) collaboratively constructed resources such as Wikipedia and (iii) Web-based resources including Web search engine results. The taxonomy of the covered knowledge resources is shown in Figure 2.

3.1.1 Linguistically constructed knowledge resources

The knowledge resources of this type consist of data sets that have been systematically developed by expert linguists. These knowledge resources are considered the most reliable as they have been curated through a well-reviewed and controlled process. Two of the most widely used resources include WordNet and GermaNet.

WordNet is a large lexical database for the English language. It consists of information that describes English words and expresses various meanings that a word can have in different contexts. Relations and synsets are two of the main constituents of WordNet where relations express information such as hypernymy, antonymy and hyponymy, and synsets represent groups of synonymous words. Additionally, members of each synset are often further described using a short piece of textual descriptor called the gloss. Various researchers have already benefited from Wordnet for computing the degree of SR between two words. These works have exploited both WordNet's relations and its glosses.

Relations in WordNet provide the means to organize words in hierarchical structures. For instance, based on the hyponymy and hypernymy relations, words can be placed in a hierarchy where relations between general and specific terms are explicitly described. This hierarchical structure expressed through WordNet relations has been the source for various SR measures through which the lowest common subsumer of two words has been used as an indication of the relation between them. Resnik (1995), Jiang and Conrath (1997) and Li *et al*. (2003) have considered the information content of the subsumer of two

https://doi.org/10.1017/S0269888917000029

words to define the degree of their relatedness. This is based on a simple yet effective observation that subsumers in lower levels of the hierarchy provide more information as they refer to more specific concepts thus revealing greater information content and hence indicating strong relatedness between two words.

Besides relying on the hyponymy and hypernymy relations, other relationship types have also been used to create a word graph where the nodes are WordNet synsets and the corresponding edges are relations. Once the graph is constructed, various graph manipulation techniques have been used to derive relatedness of the nodes in the graph. Rada *et al.* (1989) benefited from this graph representation and represented relatedness as a measure of the shortest path between two nodes. Leacock and Chodorow (1998) further improved this calculation by taking into account the depth of the graph along with the path length. In contrast and instead of focusing on path length, Hughes and Ramage (2007) applied a random walk process on the graph to extract a statistic distribution that denotes the probability of reaching other nodes by starting from a given node. SR is then computed by measuring the similarity between two static distributions obtained by starting from each of the two nodes.

While relations in WordNet allow for identifying structural relatedness between words, glosses allow for the identification of content-based relations between words. A gloss is a short piece of text that describes the meaning of each synset in Wordnet. For example, the gloss of the synset ‘relatedness’ is ‘a particular manner of connectedness’. Various notable work have already been developed that measure SR between two words based on the information content overlap of their corresponding glosses. A simple yet effective approach is to count the word overlap between two glosses, and consider the words more related if their word count overlap is higher. While Lesk (1986) introduced this method in 1986, some other methods have expanded upon it by introducing the concept of pseudo-glosses. The idea behind pseudo-glosses is that some glosses in Wordnet are too short and hence not effective for calculating relatedness. Therefore, methods are proposed to expand glosses to overcome this problem. Banerjee and Pedersen (2002) developed pseudo-glosses for a given word by concatenating the glosses of other related words (e.g. the synset, hypernym, hyponym, holonym, meronym, troponym and attribute of words in pairs) to its gloss. Mihalcea and Moldovan (1999) expanded the glosses by considering the glosses of other words in the WordNet relation hierarchy. Another approach, which deviates from the idea of word overlaps from the glosses, is based on the development of a feature representation for each word where the feature set is created using bags of words within the glosses of the words in WordNet. For instance, Patwardhan and Pedersen (2006) represented a word by its second-order gloss vector. In their work, first-order context vectors are created by measuring the co-occurrences between words based on their glosses. Then, the second-order vector for word *w* is formed by adding the first-order context vectors of words that exist in the gloss of *w*. For example, the gloss of word *fork* is ‘cutlery used to serve and eat food’, after removing stopwords, the first context vectors of words ‘cutlery’, ‘serve’, ‘eat’ and ‘food’ can be created by counting the co-occurrences between these words based on their glosses. Then the second-order vector for *fork* is created by adding the first-order context vectors of these four words.

GermaNet is a German counterpart of WordNet. Many of the approaches applied to WordNet can also be employed for GermaNet. However, the main distinguishing feature of GermaNet is that it does not include glosses; therefore, the original gloss-based methods which calculate relatedness based on glosses are not directly applicable. In order to exploit gloss-based methods, glosses need to be generated from scratch. Gurevych (2005) has proposed one such method where pseudo-glosses are generated by concatenating words that are in close relations to the target word in the relationship hierarchy. The generated pseudo-glosses are then used as a representation of the gloss for the words in GermaNet.

3.1.2 Collaboratively constructed knowledge resources

The second class of knowledge resources that are widely exploited in the literature are the information sources that have been collaboratively developed through *crowdsourcing* on the Web. While these knowledge resources are not necessarily developed by domain expert authorities, they contain reliable information due to extensive peer review and content moderation. Wikipedia and Wiktionary are among the most actively maintained information sources that have received attention from the SR community.

3.1.3 Wikipedia

The information collected in Wikipedia is represented through the so-called articles, which are focused on and dedicated to the description of a specific topic. The content of each article is gathered and edited collaboratively and is often strictly moderated by community volunteers. Besides articles, Wikipedia provides hyperlinks between articles, categories and disambiguation pages. Various researchers have already benefited from the textual content of Wikipedia articles, the hyperlink graph structure as well as categories and disambiguation pages to develop SR measures.

One of the widely used SR methods that exploits Wikipedia article content is ESA (Gabrilovich & Markovitch, 2007). In this method, each Wikipedia article is assumed to be describing a single word or concepts, which is represented as a weighted mixture of the set of terms that appear in the content of the Wikipedia article. The weights are TF–IDF values of the terms. In ESA, the main idea behind the use of Wikipedia articles is to develop a weighted bag of words representation that can be used for similarity measurement.

Article links, which are inward hyperlinks connecting two Wikipedia articles can be used to establish relationship between two concepts (words) represented by the two Wikipedia articles. Witten and Milne (2008) and Milne (2007) have already benefited from article links when proposing the WLM method. They exploit Wikipedia article links by representing each word as a weighted vector of links computed based on the number of links on the word's Wikipedia article and the probability of the link's occurrence. Different from WLM, WikiWalk (Yeh *et al*., 2009) exploits Wikipedia article link structure to construct a graph in which Wikipedia articles are the vertices and the edges are the links between the articles. This graph structure, which closely mimics the Wikipedia content structure, is employed for performing a variation of the PageRank algorithm to find word similarities.

The *Wikipedia Category* system is a hierarchical structure where each category can have subcategories through *Hyponymy* or *Meronymy* relations. Each article is coupled with one or more categories. From the category perspective, each category contains one or more articles. Given the meaningful classification that Wikipedia categories provide, WikiRelate (Strube & Ponzetto, 2006) defines SR between two words based on the mapping between the Wikipedia articles representing the words and their related categories. The basic idea behind this approach is that SR of two words is dependent on the relatedness of their categories, therefore, using the word-category mapping, the distance between the categories of two words are taken as a measure of the words' SR. Other than WikiRelate, WikiWalk (Yeh *et al*., 2009) also employs Wikipedia category links to augment the graph structure that it builds based on the article links in order to take category similarities into account.

Within Wikipedia, disambiguation pages provide context for words that can have multiple meanings. Disambiguation pages contain links to the most pertinent article per sense of the word along with a brief description. For example, querying *java* returns a Wikipedia disambiguation page which contains links to other pages consisting of Java Sea, north of the island of Java, Java Trench, a subduction zone trench west of the island of Java, among others. In addition to using Wikipedia categories, WikiRelate also benefits from the disambiguation pages by resolving all *redirects* in the disambiguation pages and selecting the sense (the redirect link) that results in the highest SR between the two words.

3.1.4 Wiktionary

Wiktionary is a multilingual, Web-based, freely available dictionary, thesaurus and phrasebook (Zesch *et al*., 2008) designed as a lexical companion to Wikipedia. Wiktionary shares many commonalities with Wordnet as they both include words, lexical relations between words and short pieces of text describing the words (glosses). Given the fact that Wiktionary consists of a large number of words, a high-dimensional concept vector can be constructed based on its constituent words. For example, Zesch *et al*. (2008) use both English and German versions of Wiktionary to compute SR. In their approach, they construct a concept vector $\underline{v}(w) = \left(v_{1,\ldots}, v_n\right)$ where the value of v_i is the TF–IDF of word w in Wiktionary entry d_i. Once each word is represented as a concept vector, SR between two words is calculated based on the cosine similarity of their concept vectors.

Similar to Wordnet, Wiktionary consist of lexical–semantic relations that are explicitly encoded in the structure of each Wiktionary entry. The English Wiktionary consists of relations such as

https://doi.org/10.1017/S0269888917000029

compounds, abbreviations and acronyms, among others (Zesch, 2010). Some researchers have developed SR measures that focus on these relations. As mentioned earlier, the work by Zesch *et al.* (2008) adopts two methods based on Wiktionary content: the first method takes Wiktionary words into account as outlined above and the second method relies on the relations between the words in Wiktionary. In the latter approach, a graph is built whose nodes are the Wiktionary words and the edges are the lexical–semantic relations between these words. SR is then measured by calculating the shortest path between each two node. Likewise, Krizhanovsky and Lin (2009) have applied a path-based method on a graph constructed based on Russian Wiktionary. In order to address the small vocabulary size of the Russian Wiktionary, the authors have used translations between the Russian and English Wiktionary. On this basis, the shortest path between two words is found and the distance is used to indicate similarity. It is also worth mentioning that Wiktionary has glosses for some of its entries. Therefore, the concept of glosses or more specifically pseudo-glosses can also be exploited for identifying SR based on Wiktionary. For example, Meyer and Gurevych (2012) explored the glosses in Wiktionary to perform disambiguation based on word overlaps in glosses. They calculated similarity between words with the right sense to create sense-disambiguated word vectors, which resulted in a higher accuracy compared to methods based on WordNet and Wikipedia.

3.1.5 Web-based resources

It has been estimated that there are over 45 billion Web pages on the World Wide Web that have been created with no central coordination[2]. Most of these Web pages carry implicit user-understandable semantics. Many researchers have relied on this implicit semantics to measure SR between words. In the Web-based knowledge resource category, two main information sources have been used, namely Web search engines and semantic Web resources.

3.1.6 Web search engines

Given the size of the Web and the role of search engines in content retrieval, there have been extensive research that have looked at how the results of search engines can be taken as an indication for SR. For a given search query, search engines often return useful information such as result snippets, Web page URIs, user-specified metadata and descriptive page titles. The information content value of the outputs of search engines have been considered as possible indicators of relatedness.

Web search engine snippets are short pieces of text for each result returned by search engine that contain a set of terms that describe the retrieved page. Some authors have benefited from snippets to measure SR. For instance, Spanakis *et al.* (2009) have proposed a hybrid Web-based measure for computing SR between words by automatically extracting lexico-syntactic patterns from snippets based on the idea that similar words should have similar usage patterns. Similarly, Bollegala *et al.* (2007) have developed a SR method that relies on search snippets, and considers both word counts and lexical-syntactic patterns when comparing the results of three queries w_1, w_2 and (w_1 and w_2). Sahami and Heilman (2006) collect snippets of the top ranked pages for a query and represent each query through an TF–IDF term vector of the collection of snippets. SR of two words is then computed based on the similarity of their query term vectors. Furthermore, Chen *et al.* (2006) have proposed a double-checking model to analyze snippets returned by a Web search engine, where the double-checking model is formed by a forward process which counts the total occurrences of w_2 in the top N snippets of query w_1 and a backward process which counts the total occurrences of w_1 in the top N snippets of query w_2. Duan and Zeng (2012) count the occurrences of each word and also the co-occurrence of the two words within the returned snippets and compute SR based on the obtained count frequencies.

There have been other works based on Web search engine results that do not necessarily rely on snippets only, but also consider the content of the retrieved Web pages. The main reason for this is the short length of snippets that could impact the accuracy of the SR measures. For example, Sahami and Heilman (2006), who initially considered snippets as their knowledge resource, have enhanced snippets by adding the top-k words with the highest TF–IDF value from each of the returned document to the vector

[2] `http://www.worldwidewebsize.com/`

https://doi.org/10.1017/S0269888917000029

representing each word. Duan and Zeng (2012) have also considered the retrieved documents by analyzing the number of documents where the two words occur independently and also those where the words co-occur. There are several works that operate based on a very similar approach on the retrieved documents, which can be found in (Bollegala *et al*., 2007; Cilibrasi & Vitanyi, 2007; Spanakis *et al*., 2009).

3.1.7 Semantic Web

More recent knowledge resources are provided by the Semantic Web community in the form of ontologies and the Linked Open Data. These resources are based primarily on the RDF model, built of triples in the form of <subject, predicate, object>. A triple explicitly defines a relationship between a subject and an object through a meaningful relationship, known as a predicate. As introduced earlier, REWOrD (Pirró, 2012) is one of the earlier works that exploit the concept of Linked Open Data, especially the DBpedia knowledge base, to compute SR. In REWOrD approach, the correspondence between words and DBpedia's semantic concepts are first found. The retrieved DBpedia concepts are then used to construct a vector for each word. Vector similarity is used as the measure of SR between two words. Gracia and Mena (2008) have calculated SR between two concepts within a Semantic Web ontology by finding and comparing the similarity of their ontological contexts. An ontological context for a concept is defined as a collection of highly related concepts within the ontology that can support unambiguous definition of the given concept. For instance, the ontological context of a concept can include its hypernyms and synonyms. Karanastasi and Christodoulakis (2007) have introduced OntoNL SR measure that depends on semantic relations defined by the Web Ontology Language. In this model, the authors compute SR by integrating three aspects: the number of common properties and inverse of properties that the two concepts share, the path distances of two concepts' common subsumer and the count of the common nouns and synonyms from the concepts' descriptions in the ontology. Finally, Zhou *et al*. (2012) have proposed the Linked Open Data Description Overlap (LODDO) method that measures SR between words as long as the words have corresponding concepts (entities) on the Linked Open Data. For any given pair of concepts, LODDO would retrieve the description of the concepts from the Linked Open Data cloud and use text overlap methods to compute the relatedness of the two concepts based on their derived descriptions.

One of the research areas that has important synergies with the SR domain is ontology matching, also known as ontology alignment. Ontology matching is concerned with the identification of similar concepts within two different ontologies. As described by Euzenat and Shvaiko (2013), and Otero-Cerdeira *et al*. (2015), ontology matching methods can be performed at element and/or structure levels. Within the element-level mapping approaches that have the closest resemblance to SR measurement techniques, there are typically two types of approaches, namely point-to-point and interlingua approaches (Gruninger & Kopena, 2005). In the point-to-point approach, elements of the two ontologies are directly mapped to each other, while in the interlingua approach, elements of each ontology are separately mapped to an intermediary standalone ontology. The idea of interlingua-based mapping can be helpful for measuring cross-language SR. It should be noted that while many mapping techniques employ SR measurement to determine the relevance or similarity of two concepts across the two ontologies, ontology mapping techniques often do not solely rely on SR and incorporate other additional information such as ontology structure and depth information to perform the mapping process.

3.2 Computational methods

In addition to the knowledge resources used for computing SR values, the method that is applied to the adopted knowledge resource plays a significant role in the quality of a relatedness measure. Methods developed for SR computation are introduced in this section. The taxonomy of such methods is shown in Figure 3. We review three major categories of methods, namely graph-based, context-based and temporal.

3.2.1 Graph-based methods

The basic idea of graph-based methods is to view the information derived from a knowledge resource as a graph whose nodes are terms or concepts, whereas edges are relations, specific to the selected knowledge resource, between pairs of terms or concepts. By adopting a graph-based representation model, many

https://doi.org/10.1017/S0269888917000029

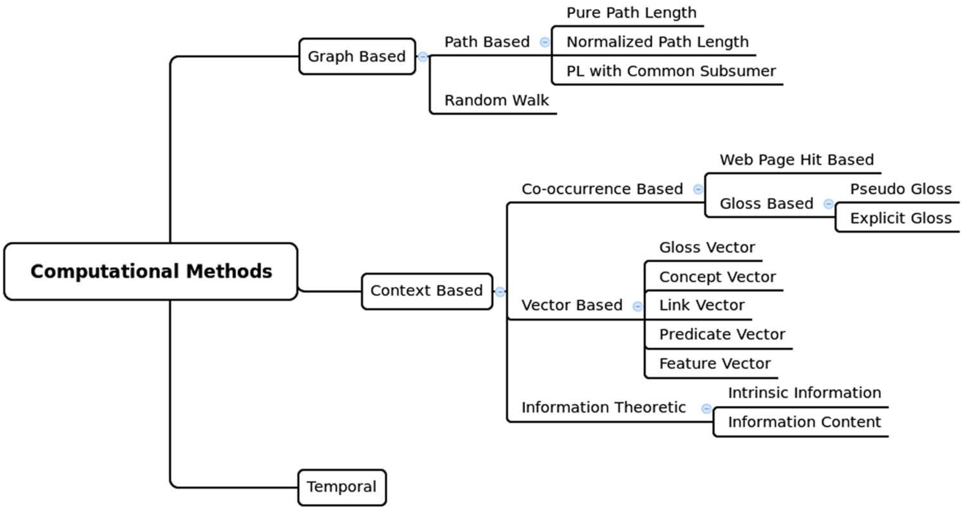

Figure 3 Computational methods taxonomy

different graph analysis techniques can be applied to compute SR between two words. The two main approaches that have been studied in the literature include *path-based methods* that consider the path length between two nodes in the graph, and *random walk methods* that take advantage of the probabilistic likelihood of reaching a destination node from a source node in the graph.

3.2.2 Path-based methods

The path connecting two nodes in a graph-based representation of a knowledge resource can reveal important information about the degree of relatedness between two words. Path-based approaches often rely on the length of the shortest path between two nodes in the graph to measure their SR. It is intuitively assumed that the shorter the path is, the higher SR between the two words would be.

Pure path length methods only consider the length of the shortest path between two nodes, which is computed by simply counting the number of edges on the path from one node to the other. For instance, Rada *et al.* (1989) compute SR by using the path length l between two nodes where the degree of similarity of two nodes is defined as the length of longest path in the graph subtracted by l. Jarmasz and Szpakowicz (2012a, 2012b) adopt Rada *et al.*'s method and apply it to Roget's thesaurus by counting the number of edges between the two words in the Roget's taxonomy. In WikiRelate, Strube and Ponzetto (2006) select the shortest path between two words (corresponding to Wikipedia articles) based on the graph constructed from Wikipedia where the nodes are the Wikipedia articles and the edges are the links between the articles. Furthermore, some researchers have used additional corpus statistics in combination with path length to compute SR. For example, Jiang and Conrath (1997) calculated the sum of all the weights on the shortest path to measure similarity, where the weights on the edges are generated from the corpus statistics.

Normalized path length approaches consider additional graph statistics such as the depth of the graph to normalize the length of the shortest path between two nodes. For instance, Leacock and Chodorow (1998) normalize the shortest path length by considering the *depth* as length of the longest path in the graph, and define the semantic similarity measure as $sim_{LC98}(w, w_2) = -\log \frac{l(w_1,w_2)+1}{2\cdot\text{depth}}$.

Both pure path length-based and normalized path length-based methods do not consider the information content value of a node. Some researchers have argued that the shared information content value of two words within a graph can be understood through their common subsumer. The consideration of the common subsumer in such approaches ensures that those words which are located higher in the taxonomy (i.e. are more abstract), receive a lower relatedness score compared to those words that are lower in the taxonomy but have comparable path length. For instance, assuming a taxonomic structure, the work by Wu and Palmer (1994) is a path length approach, which considers the lowest common subsumer of two words $lcs(w, w_2)$ along with the shortest path length between the two words in order to measure SR as follows: $sim_{WuP94} = \frac{2\cdot\text{depth}(lcs)}{l(w,lcs)+l(w_2,lcs)+2\cdot\text{depth}(lcs)}$.

https://doi.org/10.1017/S0269888917000029

3.2.3 *Random walk methods*

Some researchers have based their SR calculation on the likelihood of reaching a node from another node based on a random Markov chain traversal of the graph. In such models, the edges of the graph form a transition matrix between the vertices where each column contains a normalized outgoing probability distribution for a particular node, and the value in each cell represents the conditional probability of moving from one node to the other. Based on this initial transition matrix and with repeated conditional transitions, a stationary distribution will be obtained for each vertex of the graph. SR is computed by computing the similarity between the stationary distributions obtained for two words. For example, by extracting information from WordNet, Hughes and Ramage (2007) construct a graph where the nodes are *Synsets*, *TokenPOS* and *Tokens,* whereas the edges are the WordNet's relationships between these nodes. The authors define the probability of reaching word w_i in the tth iteration $\left(w_i^{(t)}\right)$ as the sum of all paths in the graph leading to this word from the previous iteration: $w_i^{(t)} = \sum_{w_j \in v} w_j^{(t-1)} P(w_i \mid w_j)$. Yeh *et al.* (2009) have applied random walks on Wikipedia link structure. These authors treated the Wikipedia articles as vertices and links between articles as edges of a graph. Based on this graph structure, the initial edge weights were determined based on the ESA method (Gabrilovich & Markovitch, 2007), after which the Markov chain theory was applied to obtain stationary distributions for each word. SR was then obtained by computing the similarity between any two stationary distributions.

3.2.4 *Context-based methods*

The latent relation hypothesis postulates that words that are observed in similar contexts or frequently share similar contexts can be considered related (Turney & Pantel, 2010). Context-based methods primarily operate based on this hypothesis and attempt to measure SR of words through the degree of similarity of the contexts the words appear in. Different researchers have come up with various forms of word context including Web pages where a word appears in, the Wikipedia articles where a word occurs and the WordNet glosses where the word is observed, just to name a few. We identify and elaborate on three forms of context-based SR methods, namely co-occurrence-based, vector-based and information content-based methods.

3.2.5 *Co-occurrence-based methods*

Two word contexts that have been commonly used in the literature for this purpose include (i) Web pages where the words occur, and (ii) WordNet glosses where the words are observed.

In order to exploit the Web page content where the words occur, the work proposed in Bollegala *et al.* (2007), Spanakis *et al.* (2009), Duan and Zeng (2012) employ a Web search engine to retrieve the specific Web pages where the words occur independently and also simultaneously. The degree of overlap between the retrieved Web pages for each query is used to determine relatedness. Assuming N is the number of documents indexed by the search engine and $H(q)$ is the number of search results for query q, well-known set similarity measures such as Jaccard $\left(\frac{H(w_1 \cap w_2)}{H(w_1)+H(w_2)-H(w_1 \cap w_2)}\right)$, overlap $\left(\frac{H(w_1 \cap w_2)}{\min(H(w_1),H(w_2))}\right)$, Dice $\left(\frac{2H(w_1 \cap w_2)}{H(w_1)+H(w_2)}\right)$ and point-wise mutual information $\left(\log_2\left(\frac{\frac{H(w_1 \cap w_2)}{N}}{\frac{H(w_1)}{N} \cdot \frac{H(w_2)}{N}}\right)\right)$ are used to measure SR of two words w_1 and w_2.

A seminal work in this area is the Google distance proposed by Cilibrasi and Vitányi (2007). The authors have proposed the NGD to determine the distance between a pair of words where the degree of relatedness is determined based on Google's search results. If two words produce the exact same search result set when used as a query in the Google search engine, their NGD would be 0 and if they do not share overlaps, their NGD would be infinite. Gracia and Mena (2008) later transformed NGD to compute the relatedness between words regardless of whether Google search engine is used or not.

As mentioned earlier, context has also been modeled through WordNet glosses where each word's gloss or any gloss where the word is observed are considered to be the context for the word. Many of the existing work such as Lesk (1986), Gurevych (2005), Zesch *et al.* (2008) are based on such context

https://doi.org/10.1017/S0269888917000029

definition and assume that each word has either a WordNet entry with a corresponding gloss or a gloss can be synthetically generated for the word.

When context is modeled as through explicit glosses, they are extracted directly from WordNet. For example, Lesk (1986) built his method by counting the number of word overlaps between two glosses: $|\text{gloss}(w_1) \cap \text{gloss}(w_2)|$, where gloss($w_i$) is the set of words in the gloss of word w_i. Banerjee and Pedersen (2002) extended the gloss of each word by taking into account the glosses of related words to overcome the problem that some glosses in WordNet are short in length. Moreover, Mihalcea and Moldovan (1999) constructed the gloss of a word by combining all the glosses found in its synsets, and then counted the number of word overlaps to determine relatedness.

Considering explicit glosses and their extensions as word context is not always possible, for example, the case of GermanNet; therefore, in some cases pseudo-glosses are employed as context. For instance, Gurevych (2005) constructed pseudo-glosses by concatenating words which are in close relation (e.g. Synonymy, Meronymy) with the target word.

3.2.6 Vector-based methods

The idea behind vector-based models is to construct a vector representation model for each word that can be used to calculate SR through vector similarity measures. Word vectors have been traditionally represented using information extracted from different knowledge resources such as WordNet glosses, Wikipedia links and Web search result snippets, just to name a few. Based on the type of elements used in the word vector representation, we divide vector-based methods into gloss vector, concept vector, link vector, predicate vector and feature vector categories.

Within the *gloss vector* category, Patwardhan and Pedersen (2006) construct word vectors using WordNet glosses. The authors initially create the first-order co-occurrence vectors in WordNet, where the co-occurrences are between the target word and other words in the target word's gloss. This is followed by computing second-order co-occurrences, which is inspired by the *second-order word sense discrimination* approach proposed by Schütze (1998). The authors suggest that the use of the Cosine similarity measure on any two such vectors would result in a reliable SR value for those two words. Other researchers have later proposed some variants of the gloss vector representation such as the work by Wan and Angryk (2007) and Pedersen (2012).

While gloss vector methods focus on information from WordNet, *concept vector* methods employ content from Wikipedia to build vector representation of words. One of the better known concept vector method, introduced by Gabrilovich and Markovitch (2007), is based on the assumption that each Wikipedia article has a *topical* focus, that is, the content of each Wikipedia article is focused on a specific topic. Accordingly, a word is represented as a vector whose elements are the TF–IDF values of the words that appear in the corresponding Wikipedia article. The limitation of this approach is that it only provides SR values between words that have corresponding Wikipedia articles. Zesch *et al.* (2008) also created a high-dimensional concept vector for each word based on the concept space in Wiktionary.

Unlike gloss and concept vector models, *link vector* methods represent a word through its links to other words. For this purpose, the link vector model needs to be built on knowledge resources that provide some form of word interlinking, for example, through hyperlinks. Milne (2007) has proposed one of the widely used link vector models where each word is represented by the links it has to other Wikipedia articles. However, given the fact that not all the links in an article have the same significance, the author defines a weighting scheme for the links. The basis of the weighting scheme is that a page would be considered rather general (less specific) if too many pages link to it. Therefore, Milne defines the weight of a link in a specific Wikipedia article as $\log\left(\frac{N}{|T|}\right)$, where T is the number of articles that link to the target article and N the total number of Wikipedia articles. A word is then expressed as a weighted vector of links that appear in its corresponding Wikipedia article. Other authors such as Bu *et al.* (2011) and Turdakov and Velikhov (2008), among others, have also used and promoted the link vector representation.

In the *predicate vector* representation, the focus is on deriving a vector for each word based on the content of RDF graphs. For instance, in the REWOrD system, Pirró (2012) created a predicate vector for each word, in which the elements of the vector were other words that were connected to the target word through at least one explicit predicate. The author further suggested that the predicate vector could contain other words that were observed along the path of the words that were compared for SR. Predicate frequency, inverse triple frequency and path informativeness were used to weight each element of the vector.

Finally, *feature vector* models focus on identifying key discriminative characteristics that can uniquely represent a word. The major difference between feature vector models and the previous three vector representations is that the elements of the feature vector do not rely directly on some form of co-occurrence information but rather they rely on specific metrics to represent a word. For instance, Spanakis *et al*. (2009) proposed to model each word as a feature vector that includes features such as page count metrics, and lexico-syntactic patterns extracted from Google results (e.g. using titles, snippets and URLs). Along the same lines, Bollegala *et al*. (2007) constructed a feature vector based on the lexico-syntactic patterns that they extracted from the results of a Web search, for example, a word pair based on the frequency of each pattern. For instance, they determined that words that were related to each other in a given sentence using phrases such as: *also known as*, *is a*, *part of*, *is an example of*, have a high likelihood of being suitable features for computing SR.

3.2.7 Information theoretic methods

Information theoretic approaches compute relatedness between words by considering how much common information the two words share. The intuition is that the more similar information the two words convey, the more similar they would be. Information theoretic approaches can be divided into two subcategories depending on how information content sharing is measured.

Intrinsic information theoretic methods rely on a *taxonomic* knowledge resource for measuring SR. To determine the degree of common information shared by two words, intrinsic methods consider features such as position and frequency of the word in the taxonomic structure. For instance, Resnik (1995) proposed one of the seminal intrinsic methods where similarity of two words is determined by considering the information content value of two words' subsumer. In his work, Resnik defines information content as the negative log likelihood of the probability of encountering an instance of a given concept. In simple terms, the more general the common subsumer of the words is in the taxonomy hierarchy, the less similar the words would be. Later, Seco *et al*. (2004) base their work on the primary premise of Resnik's work by assuming that infrequent words are more informative than frequent ones. Based on this assumption, information content value of a word is determined within the context of WordNet by counting the number of hyponyms that a word has, where the words that have more hyponyms have a lower information content value. Furthermore, the authors assume that words that are leaf nodes in the WordNet hierarchy can be assumed to exert maximal information content.

In the other class of information theoretic approaches, known as *information content methods*, the information value of the words is considered for computing SR. Among the better known works in this class, Jiang and Conrath (1997) and Li *et al*. (2003) have extended the approach developed by Resnik (1995) by additionally making use of the information content of a word. In Jiang and Conrath's (1997) work, two measurements were used, namely, node-based information content calculation and edge counting. In the node-based approach, the information content of a concept in a taxonomy is defined as the probability that an instance of that concept is encountered in that taxonomy. In the edge counting schema, distance is calculated between two nodes representing instances of the concepts being compared. The shorter the distance is, the more similar the two concepts are. Jiang and Conrath found that the edge counting scheme is highly dependent on the quality of the taxonomy and its structure while the node-based approach is less sensitive to the details of the hierarchy of the taxonomy. The authors further proposed an edge-based approach where the distance function between two concepts is defined as the sum of two concepts' information content subtracted from the information content of the concepts' lowest super-ordinate. Furthermore, Li *et al*. (2003) works with a tree-structured taxonomy and intuitively assumes that: (1) similarity between two concepts is related to their commonality, where commonality is measured by the number of nodes in the taxonomy that related to both concepts; (2) similarity between two concepts is also dependent on the differences between them where the difference between two concepts is measured by the number of nodes that exclusively

https://doi.org/10.1017/S0269888917000029

related to each concept but not the other; and (3) the maximum similarity between two concepts is when they are identical. Lin defines the information content of a concept based on the probability that randomly selected nodes in that taxonomy belong to that concept. Accordingly, SR is measured based on the information content of similarity and differences between two concepts.

3.2.8 Temporal methods

Some researchers have recently focused on the temporal correlation between words to determine their SR. While there are not many approaches that consider temporality, the idea behind such approaches is that words that have similar behavioral patterns over time, for example, occurrence, can be considered similar. Temporal methods require knowledge resources that incorporate and offer some form of temporality in their information. For instance, Radinsky *et al.* (2011) propose the TSA method where they represent each word as a weighted vector of word time series produced from a historical archive such as the history of Wikipedia articles, which shows the temporal evolution of each article. Based on the time series for each word, the SR of two words is measured through time series cross-correlation and dynamic time warping. In a different line of work, Milikic *et al.* (2011) have been one of the earlier researchers who have used a non-traditional knowledge resource for temporally modeling SR. These authors measured co-occurrence of words on Twitter to calculate SR of those words; then, standard deviation of the SR of words within different time periods is employed to estimate the changes of SR between words over time. Finally, Zhao *et al.* (2006) hypothesized that temporal factors have a strong impact on the accuracy of similarity measures especially in the context of search queries. Therefore, the authors present a framework that considers temporal characteristics of historical search click-through data to enhance the measurement of similarity between queries. In their work, the similarity between search queries is determined based on the similarity of their historical click-through pages over several different time periods.

3.3 Evaluation

In order to evaluate their SR work, researchers have used different data sets and methods for comparative analysis. In this section, we focus on classifying the *data sets* and *methods* that exist in the literature for evaluating SR methods.

The data sets that have been used in the evaluations are mostly curated for the English and German languages. These data sets are often constructed by collecting subjective opinion of humans with regards to the SR of words. Table 2 provides an overview of some of the common data sets and their brief description. From the perspective of the evaluation methods, these methods can be divided into two main categories, namely determining correlation with human judgments and application-specific evaluations. Table 3 provides a summary of the evaluation methods that have been used in the literature. The taxonomy of the data sets and methods used for evaluating SR methods is shown in Figure 4.

3.3.1 Data sets

The main purpose of developing SR data sets is to curate a set of word pairs with known degrees of SR so they can be used as a gold standard benchmark for evaluating various SR methods. The data sets are most often developed by soliciting human judgments with regards to the SR of a set of word pairs. The data sets that have been used and cited in the literature are primarily in the English and German languages.

As shown in Table 2, the most popular English language data sets are the RG-65, MC-30, Fin-353 and YP-130 data sets:

- The Rubenstein and Goodenough (1965) data set (RG-65) includes 65 noun pairs. The similarity of each pair is scored on a scale of 0–4 where a higher number indicates higher similarity. In order to collect human judgments, 51 subjects participated in the data collection process and the similarity value of each word pair is equal to the average of the values assigned by the subjects. The RG-65 data set has been used by many researchers as a gold standard to evaluate their SR methods, for example, Strube and Ponzetto (2006) and Gabrilovich and Markovitch (2007) selected RG-65 as a gold standard to analyze their work.

https://doi.org/10.1017/S0269888917000029

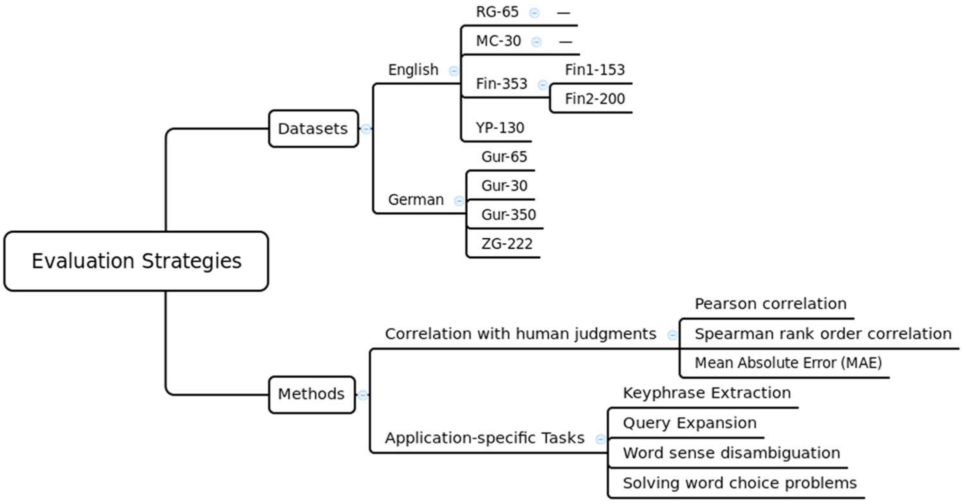

Figure 4 The details of the evaluation strategies

- The Miller and Charles (1991) data set (MC-30) is a subset of 30 pairs taken from the original RG-65 data set with an additional replicated experiment by another 38 subjects. Given the replicated study and a relatively manageable size of word pairs, the MC-30 data set has been one of popular data sets for comparative analytics in many works such as Witten and Milne (2008) and Spanakis *et al.* (2009).
- The Finkelstein *et al.* (2002) data set (Fin-353) contains 353 English word pairs among which 30 word pairs are directly taken from the MC-30 data set. The data set is further divided into two subsets where the scores in first set, Fin-153 (containing 153 word pairs), are obtained from 13 subjects, and in the second set, Fin-200 (containing 200 word pairs), from 16 subjects. Therefore, in some works, the first set has been used for training purposes, and the second set is then used for evaluation. The use of Fin-353 has also been quite popular in the literature including the work by Pirró (2012) and Agirre *et al.* (2009), among others.
- The Yang and Powers (2006) data set (YP-130) contains 130 verb pairs particularly made for evaluating the ability of a SR method to determine the relatedness of verbs. Zesch *et al.* (2008) are one of the few researchers that employed the YP-130 data set in order to evaluate the ability of their proposed SR on verb pairs in addition to more typical noun pairs.

Researchers have also developed data sets in German among which Gur-65, Gur-30, Gur-350 and ZG-222 are the most popular:

- The Gurevych (2005) data set (Gur-65) is the German translation of the English RG-65 data set. Gurevych (2005) and Zesch *et al.* (2008) have used the Gur-65 data set to evaluate their methods.
- The Gurevych (2005) data set (Gur-30) is a subset of the Gur-65 data set that corresponds to the English MC-30 derived from RG-65.
- The Gurevych (2006) data set (Gur-350) contains 350 word pairs which includes nouns, verbs and adjectives curated by eight human subjects. Although used only in few works, such as Zesch *et al.* (2008), it is a valuable data set that includes a wide variety of word types that cannot be seen in other data sets.
- The Zesch and Gurevych (2006) data set (ZG-222) consists of word pairs from specific domains. It contains 222 domain specific word pairs that were evaluated by 21 subjects. This data set also consists of nouns, verbs and adjectives.

3.3.2 Methods

The typical methods for evaluating SR techniques can be broadly classified into two classes: (1) computing the degree of correlation with human judgments, and (2) measuring performance in application-specific tasks.

https://doi.org/10.1017/S0269888917000029

Table 2 Summary of the data sets reported in the literature

Data set	Language	Citation	Number of subjects	Description	SR methods
RG-65	English	Rubenstein and Goodenough (1965)	51	Includes 65 noun word pairs with scores from 0 to 4	Patwardhan and Pedersen (2006), Strube and Ponzetto (2006), Hughes and Ramage (2007), Witten and Milne (2008), Zesch *et al.* (2008), Pirró (2012)
MC-30	English	Miller and Charles (1991)	38	30 pairs taken from the original RG-65 data sets	Resnik (1995), Jiang and Conrath (1997), Patwardhan and Pedersen (2006), Strube and Ponzetto (2006), Bollegala *et al.* (2007), Hughes and Ramage (2007), Witten and Milne (2008), Zesch *et al.* (2008), Spanakis *et al.* (2009), Yeh *et al.* (2009), Pirró (2012)
Fin-353	English	Finkelstein *et al.* (2002)	16	Contains 353 English word pairs where 30 word pairs are from MC-30	Strube and Ponzetto (2006), Gabrilovich and Markovitch (2007), Hughes and Ramage (2007), Milne (2007), Witten and Milne (2008), Zesch *et al.* (2008), Spanakis *et al.* (2009), Yeh *et al.* (2009), Radinsky *et al.* (2011), Duan and Zeng (2012), Pirró (2012)
YP-130	English	Yang and Powers (2006)	6	Contains 130 verb pairs	Zesch *et al.* (2008),Taieb *et al.* (2013)
Gur-65	German	Gurevych (2005)	24	German translations of the English RG-65	Gurevych and Niederlich (2005), Zesch *et al.* (2007, 2008)
Gur-350	German	Gurevych (2006)	8	Contains 350 German word pairs	Zesch *et al.* (2008)

SR = semantic relatedness.

3.3.3 Correlation with human judgments

One of the main techniques for evaluating SR methods has been to compare their outcomes with a gold standard data set such as those introduced earlier. Researchers have either compared the absolute predicted relatedness value with the relatedness value of the gold standard, or compared the word pair rankings produced by the relatedness method with the rankings in the gold standard. The latter approach has received more reception as it is less sensitive to the actual relatedness score values and allows for a more pragmatic comparison of the relatedness measures in practice. Such an approach hypothesizes that in order to be considered an accurate SR method, the produced rankings from the word pair orderings need to be accurate regardless of the actual numerical value assigned to word pairs. However, in the former evaluation method, the absolute SR values are considered to be important with the justification that the rankings in the gold standard data sets do not necessarily accurately represent the desired word pair ordering. This is supported by the fact that in some cases, the gold standard word orderings are sensitive to very small difference between the word pair similarities and therefore, the correct order is questionable.

https://doi.org/10.1017/S0269888917000029

Table 3 Summary of methods used for evaluation of semantic relatedness (SR) methods

Type	Method	SR methods
Measurement based on human subject gold standard	Pearson's correlation	Strube and Ponzetto (2006)
	Spearman's correlation	Patwardhan and Pedersen (2006), Gabrilovich and Markovitch (2007), Hughes and Ramage (2007), Milne (2007), Gracia and Mena (2008), Witten and Milne (2008), Yeh *et al.* (2009), Radinsky *et al.* (2011), Duan and Zeng (2012), Pirró (2012)
	MAE	Ferrara and Tasso (2013), Feng *et al.* (2015)
Measurement based on application level evaluation	Query suggestion	Vélez *et al.* (1997), Sahami and Heilman (2006)
	Community mining	Chen *et al.* (2006), Bollegala *et al.* (2007), Matsuo *et al.* (2007), Mika (2007)
	Entity disambiguation	Schütze (1998), Bollegala *et al.* (2006), Bollegala *et al.* (2007)
	Solving word choice problems	Turney (2006), Zesch *et al.* (2008), Jarmasz and Szpakowicz (2012a, 2012b)
	Word sense disambiguation	Resnik (1999), Patwardhan and Pedersen (2006), Sabou *et al.* (2007), Gracia and Mena (2008)
	Ontology matching	Sabou *et al.* (2007), Gracia and Mena (2008)
	Keyphrase extraction	Mori *et al.* (2007), Zesch (2010)

MAE, mean absolute error.

In order to evaluate the absolute value of the predicted SR measure, researchers have predominantly used the mean absolute error which measures how closely the predicted value resembles the expected value (Polčicová & Návrat, 2002; Bicici, 2015). For the purpose of measuring rank correlations, Spearman's rank correlation has been used (Gabrilovich & Markovitch, 2007; Hughes & Ramage, 2007). Spearman's correlation compares if the ranking of the results from a specific SR method correlate with the ranking provided by human judgments in the gold standard. Pearson's product–moment correlation has also been used by some researchers such as Strube and Ponzetto (2006)

3.3.4 Application-specific tasks

As an alternative to the direct evaluation of SR methods through a gold standard, application-specific tasks are often used to measure the impact of the proposed SR methods on improving the performance of a particular task. The underlying hypothesis of application-specific evaluations is that the more accurate a SR measure is, the more it improves the performance of the task at hand. Different authors have used various application-specific tasks for evaluating their work. For instance, Sahami and Heilman (2006) evaluated their work in the context of search query suggestion; Bollegala *et al.* (2007) considered the community mining domain to test their SR method; Zesch *et al.* (2008), Patwardhan and Pedersen (2006) and Gracia and Mena (2008) considered entity and word sense disambiguation as their target evaluation application area; Gracia and Mena (2008) deployed their method in the context of the ontology matching task.

The advantage of application-specific tasks-based evaluation is that not only it shows whether the SR measure is able to cause any notable improvement but also shows how well the SR measure is suitable for domain specific tasks. For instance, one could show, through experimentation, that although a given SR method does not perform well under all conditions, it is effective for a specific task or application area. Table 3 shows how different work in the literature have implemented and reported their evaluation strategy and results.

4 Comparison of the different methods in framework

In this section, we map the selected methods into the proposed framework. To this end, we have created three mapping tables based on the three top-level dimensions in the framework: Knowledge Resources

Table 4 Summary of used knowledge resource

	Knowledge resource						
	Linguistically constructed			Collaboratively constructed		Web based	
System	WordNet	GermaNet	Others	Wikipedia	Wiktionary	Search Engine	Semantic Web
Resnik	✓						
Jiang and Conrath	✓						
Lesk	✓						
ESA				✓			
Cilibrasi and Vitányi						✓	
WikiRelate!				✓			
Sahami and Heilman						✓	
Patwardhan and Pedersen	✓						
Hughes and Ramage	✓						
TSA				✓			
WLM				✓			
Zesch *et al.*					✓		
Gur		✓					
REWOrD							✓

ESA = Explicit Semantic Analysis; TSA = Temporal Semantic Analysis.

(Table 4), Methods (Table 5) and Evaluation Strategies (Table 7). In these tables, the columns show the dimensions and sub-dimensions of our framework and the rows are the methods studied here, and each cell presents the value of the dimension for the selected method.

In order to help researchers or system builders develop their SR methods by selecting different features according to their requirements, we summarize the differences, advantages and weaknesses of each dimension in the framework.

4.1 Selection of knowledge resources

The knowledge resource selected as the underlying foundation for computing SR defines primarily how the relationship between the words is viewed. Linguistically constructed knowledge resources accurately model the relations between words and provide reliable definitions of words given they are most often constructed by expert linguists. However, accurate construction of such knowledge resources is expensive and time consuming and as new words are continuously being added to the language, it is becoming increasingly hard to maintain such resources. Still, majority of the SR methods covered in this paper, use linguistically constructed knowledge resources due to their accuracy and reliability.

Collaboratively constructed knowledge resources, such as Wikipedia, are created through crowdsourcing. In Wikipedia, articles provide tremendous amount of information about contexts where certain words appear, the co-occurrence patterns, link structure of content relationships, word sense variants, and even word and concept categories, which have all been gathered through crowdsourcing. The collaborative nature of such knowledge resources enables efficient and continuous update of information; therefore, new additions to the language are more likely to be covered. According to a report from Zesch and Gurevych (2010), the growth of Wikipedia has a positive effect on coverage without affecting the suitability and accuracy of results. Another unique characteristic of collaboratively constructed knowledge resources is that the involvement of many authors leads to the incorporation of many different distinct styles of writing and word selection, which while may not be ideal for the coherency of the text itself, is an ideal source of information about people's tendency toward word usage and word relatedness.

Table 5 Summary of the discussed methods

	Computational methods														
	Graph based				Context based										
	Path based				Co-occurrence based			Vector based					Information theoretic		
						Gloss based									
System	Pure path length	Normalize path length	Path length with common subsumer	Random walk	Web page hit based	Pseudo-gloss	Explicit gloss	Gloss vector	Concept vector	Link vector	Predicate vector	Feature vector	Intrinsic information	Information content	Temporal
Resnik													✓		
Jiang and Conrath														✓	
Lesk							✓								
ESA									✓						
Cilibrasi and Vitányi					✓										
WikiRelate!	✓														
Sahami and Heilman												✓			
Patwardhan and Pedersen								✓							
Hughes and Ramage				✓											
TSA															✓
WLM										✓					
Zesch *et al*.									✓						
Gur						✓									
REWOrD											✓				

ESA = Explicit Semantic Analysis; TSA = Temporal Semantic Analysis.

While both linguistically and collaboratively developed knowledge resources provide descriptive information about words, other sources of textual content such as those provided through the Web in general, for example, Weblogs, news outlets and social networks, can be used as an informal source of word semantics. Our recent work showed that the semantics of words might shift depending on the context where they are used or where they appear (Feng *et al.*, 2015). For instance, there seems to be an observable difference in the most common senses of words when used on Twitter compared to when the words are used on Wikipedia. For this reason, Web content, retrievable through Web search engines, can provide a valuable source of information about word semantics based on their occurrence contexts. However, while this type of resource provides a very high coverage, the accuracy of the information is dependent on the quality of the search engine and the degree of ambiguity of the terms that being queried. Table 4 summarizes the use of various knowledge resource types by the selected methods.

4.2 Selection of computation method

The selection of the most suitable method for computing SR depends on many different factors such as the type of knowledge resource that is adopted, the amount of computing and storage resources available for the computation, and the desired accuracy of the approach, just to name a few. For instance, one would only be able to adopt a path-based method if the underlying knowledge resource can be modeled through a graph representation. Furthermore, depending on the type of the path-based method, the explicit type of edges in the graph need to also be well defined, for example, in the case of those methods that rely on the common subsumer of two nodes, the type of edges connecting two nodes is the subsumption relation.

Unlike path-based methods, random walk-based methods do not require explicit semantics of the relations to be defined in a knowledge resources; they only need the edges to be of the same type and have a quantifiable weight, which could for instance be the co-occurrence number of two words. Therefore, compared to path-based methods, methods that adopt a random walk approach have fewer requirements on the underlying knowledge resource and can be used in conjunction with a wider range of knowledge resources.

Context-based methods can be applied on any knowledge resource that includes minimal description of words; therefore, they are much more flexible and can be used with various types of knowledge resources. For instance, co-occurrence-based methods calculate word overlaps in textual information, which can be easily extracted from any source. However, the limitation of such approaches is that information about the various senses of a word is not directly considered and therefore there is a possibility that the usage pattern of ambiguous terms can negatively impact the accuracy of the SR scores. One of the pitfalls of the context-based approaches is the role of semantically insignificant words that appear in many different contexts. Such words co-occur with many words and therefore in many cases increase the probability of SR between two words that are otherwise not related.

Similar to context-based approaches, vector-based methods do not have specific requirements from the underlying knowledge resource. In such approaches, each word is represented as a vector of features. The most common vector representation is the bag of words model derived from different knowledge resources. When designing vector-based models two important considerations need to be taken into account: (i) the bag of words representation for words is extremely sparse and often overlooks word interdependencies. More recent approaches for the vector representation in NLP such as Word2Vec (Mikolov *et al.*, 2013a, 2013b) and deep semantic embedding (Wu *et al.*, 2014) can be used to improve this. (ii) the model is highly sensitive to the weights of words in the vector (Turney & Pantel, 2010); therefore, the decision as to which weighting scheme to be used in the vector would have a high impact on the results. The weighting schemes that require global corpus information would need more computation and update as the corpus evolves. Therefore, while quite straightforward to implement, vector-based models are quite sensitive to features used in the vector representation and the weights applied to the features.

Information theoretic methods are one of the most restricted models as they are highly coupled with the underlying knowledge resources, which need to have a structured form. The structure of the knowledge resources is used to determine the degree of information that a pair of words share. Therefore, only

https://doi.org/10.1017/S0269888917000029

Table 6 Inter-rater agreement values of each data sets

Data set	Language	InterAA
MC-30	English	0.90
YP-130	English	0.87
Gur-65	German	0.81
RG-65	English	0.80
Fin1-153	English	0.73
Gur-350	German	0.69
Fin2-200	English	0.55
ZG-222	German	0.49
Gur-30	German	—

InterAA = Inter-rater agreement.

knowledge resources such as WordNet can be used in information theoretic methods, thereby, restricting the applicability of such approaches in practice.

4.3 Selection of evaluation technique

In terms of evaluating the developed SR measures, our review shows that most authors have adopted a gold standard-based approach and compared their results with the gold standard according to the derived ranking of the word pairs using Spearman's rank correlation. As shown in Table 6, there are different gold standard data sets. One of the important factors in deciding which gold standard data set to adopt is the inter-rater agreement of the participants from whom the similarity values were collected. Table 6 reports the inter-rater agreement of the participants for the gold standard data sets where available. As argued by Graham *et al*. (2012), an inter-rater agreement of over 75% would be considered reliable; therefore, gold standard data sets with such agreement or higher can be effectively used in experiments.

One of the reasons that application-specific tasks have not been widely used in the literature is that the accuracy of the SR method is not directly observable and is only evaluated indirectly through the performance of the higher level task. Therefore, it is possible that a good performing SR method is affected by the parameters inside the application framework. In order to properly use application-specific tasks for evaluation of SR measure, a controlled experiment needs to be organized where all parameters of the application-specific task are kept constant and the SR method would be the only variable parameter. The performance of the task would then be measured and directly compared before and after the SR method is applied to the task to measure its impact.

In summary and according to Table 7, for the purpose of evaluation, most authors have chosen to work with RC-65, MC-30 and Fin-353 data sets as their gold standards, in combination with Spearman's rank correlation method.

5 Meta-analysis

From the comparative analysis of the different SR methods that have been reviewed in the previous sections, two distinct approaches for calculating semantic similarity emerge: (1) approaches based on latent relation hypothesis, and (2) approaches based on content structure.

In the approaches that adopt the latent relation hypothesis, the main premise is that relatedness is derived from and measured based on the words' context. For this reason, knowledge resources are primarily employed to build context for every given word. In these approaches regardless of the content of the knowledge resource, the primary objective is to identify a representative context for a word such that similarity and relatedness between words could be determined based on the similarity of their contexts. Context is predominantly defined as the words or terms that surround the word of interest or that are used to define it. Therefore, many different types of textual corpora such as Wikipedia articles, Web pages, search snippets, WordNet glosses, among others have been used to build context for words.

https://doi.org/10.1017/S0269888917000029

Table 7 Summary of the use of evaluation strategies

	Evaluation strategy												
	Data sets						Methods						
	English				German		Correlation with human judgments			Application-specific tasks			
System	RG-65	MC-30	Fin-353	YP-130	Gur-65	Gur-350	Pearson's correlation	Spearman's rank order correlation	Mean absolute error	Keyphrase extraction	Query expansion	Word sense disambiguation	Solving word choice problem
Resnik[1]		✓						✓					
Jiang and Conrath		✓						✓					
Lesk[2]													
ESA			✓					✓					
Cilibrasi and Vitányi													
WikiRelate!	✓	✓	✓				✓						
Sahami and Heilman											✓		
Patwardhan and Pedersen	✓	✓						✓				✓	
Hughes and Ramage	✓	✓	✓					✓					
TSA			✓					✓					
WLM			✓					✓					
Zesch *et al.*	✓	✓	✓		✓	✓		✓					✓
Gur					✓			✓					
REWOrD	✓	✓	✓					✓					

ESA = Explicit Semantic Analysis; TSA = Temporal Semantic Analysis.

[1] Authors did not specify which correlation method was used.

[2] The paper did not include an evaluation strategy.

The hypothesis is that if two words or terms are observed frequently enough together, then they are most likely related to each other from a semantic perspective, as well. The advantage of such an approach is that it is able to determine SR of words, terms or phrases that do not have explicit semantics or presence in formal dictionaries, for example, *brb*, *icymi*, among others. It has been shown by Feng *et al*. (2015) that the latent relation hypothesis is a suitable model for handling the SR of terms in contexts where unfamiliar terms are abundant such as Twitter.

While the approaches that focus on building context for words based on textual corpora have the flexibility to handle a wide range of words and terms, they can be sensitive to the role and impact of noise. The authors of ESA (Gabrilovich & Markovitch, 2007) have pointed out that they have crawled information from the Open Directory Project and used it for training purposes, which could have led to 'non-negligible amount of noise'. In general and when considering methods that employ knowledge resources that are derived from open information sources, the impact of noise needs to be measured and controlled. This is an issue that the research community is yet to explore. Within the context of topic and event detection from social network content, some authors have already proposed information theoretic and time series-based methods for identifying noise from textual corporal (Weng & Lee, 2011). Such models can be applied to the knowledge resources used in the context of SR to avoid the impact of noise on the measurements.

Another important consideration when dealing with large corpora to build context is the domain specificity of the content and its impact on the dominant senses of the words. For instance, depending on the domain to which the corpora belongs, the set of words that co-occur with ambiguous terms such as apple, and java would be different. For these two examples, if the domain of the corpora being used is technology, then the inclination of the terms that co-occur with these two words would lean toward the Computer Science sense of the words. However, if the corpora being used is on food, then the other sense will dominate. The SR derived based on one corpus would therefore not be transferrable to the other. For this reason, it is imperative to systematically understand the role of the underlying corpora and their domain on the SR measure that is built. To the extent of our knowledge, there is yet to be work on the transferability of SR method across corpora. For instance, would the SR models learnt from Wikipedia content be applicable for similarity measurement on online News content from CNN Politics?

It is also important to point out the role of temporality when context is being constructed for words. Most of the work (except a few such as Radinsky *et al*., 2011) in the literature assume that the knowledge resource that is being used as the basis for SR measurement is stable and constant. Therefore, the context that is constructed on this basis does not consider the possibility of the evolution of the word contexts over time. For instance, corpora gathered prior to 1976 will not have any reference to an entity called Apple Inc. However, such references will show up after this date. Therefore, the context of the word apple changes significantly over time and needs to be taken into consideration. The same applies to many words that have either been created after a certain date or their meaning or interpretation has changed or continues to change in reaction to real-world events. Even more so now than before, with the real-time nature of the content on the Web, the speed at which words contexts evolve is much faster; therefore, there is need for not only methods that consider temporal contexts but also methods that are able to automatically determine the length of the time windows between which the semantics of the words shift through time. The time window determination would be important as the time window for each word might be different depending on how their semantics change over time. Some words would essentially carry the same semantics over a long period of time while others might have much faster semantic shifts.

In the second category of work, knowledge resources are used to extract structure that can be used for determining SR. For instance, links between Web pages, Wikipedia category hierarchy and WordNet hypernym links have been used to build graphs from the knowledge resources to be used for measuring SR. Most existing online content forms provide some method of direct or indirect linking mechanism between content items; therefore, the extraction of structure from such corpora is possible. The added benefit of adopting structure extraction approaches from knowledge resources is that it opens up the possibility of applying a wealth of techniques that can analyze structured content, for example, tree traversal and path finding techniques, graph analysis techniques and network mining, just to name a few. As was discussed in the paper, various authors have used a variety of different knowledge resources for building the structure that they need for computing a semantic similarity measure. Hence, the SR value will

https://doi.org/10.1017/S0269888917000029

be dependent on the knowledge resource that is used to extract the structure. For instance, the SR methods of WikiRelate! That use the Wikipedia category tree structure would only be suitable for measuring the SR between two words that are observed in Wikipedia and not beyond that. To address this limitation, one of the areas that warrants further exploration is the integration of various structures that can be obtained from different knowledge resources. For instance, it would be interesting to investigate whether it is possible to integrate structures extracted from WordNet, for example, hypernym relation graph, with the graph extracted from Wikipedia category hierarchy so that a more comprehensive representation is built. Such an integration would allow the SR methods to not only have a higher coverage but also be more accurate as information from multiple facets are integrated.

In addition, and along similar lines, there may be cases where structure information or context-based information alone would not be sufficient for accurately modeling SR of words. There have been very few reported works that have explored the possibility of systematically integrating structural and context-based information. For instance, within the domain of user interest modeling on microblogging platforms such as Twitter, researchers have already explored the possibility of integrating structural information and context-based information, for example, users' social network structure combined with their posted content, for identifying user interests, which have shown to be quite effective (Zarrinkalam *et al.*, 2016). Similar work on the systematic integration of context and structure could lead to more robust SR measurement techniques.

A final observation that can be made from reviewing the literature is the role of the communication medium on determining the SR of words. As reported in Feng *et al.* (2015), the semantics of words can substantially shift depending on the communication medium where they are used. For instance, the words movie and popcorn were determined to be highly related to each other on Twitter but rated very low within the WS-353 gold standard. Therefore, it seems that SR is not only context dependent but may also be dependent on the medium where the communication takes place. An important area where further exploration can happen is to understand how medium impacts the semantics of words and whether it would be possible to align the semantic spaces of words across different communication medium.

6 Concluding remarks

In this paper, we report on a comprehensive study of SR methods, which considers different knowledge resources, methods and evaluations. First, we selected a representative set of SR approaches reported in the literature. Then, we created a framework to classify these approaches according to the following three dimension: knowledge resource(s) used, the computational method applied for computing relatedness and the adopted evaluation method. By mapping the selected systems into the framework, we systematically analyzed the advantages and disadvantages of each identified knowledge resources, relatedness computational method, as well as evaluation methods. Therefore, researchers who would want to further improve or deploy certain SR systems or methods can highly benefit from the insight provided by this study. In addition, we have provided a critical discussion on the limitations of existing methods and offer suggestions on potential valuable research directions that can be taken in future research in this domain.

Acknowledgments

The first two authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada.

References

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

Banerjee, S. & Pedersen, T. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing '02)*, Gelbukh, A. F. (ed.). Springer-Verlag, 136–145.

Bicici, M. E. 2015. RTM-DCU: predicting semantic similarity with referential translation machines. In *SemEval-2015: Semantic Evaluation Exercises – International Workshop on Semantic Evaluation*. http://doras.dcu.ie/20650/.

Bollegala, D., Matsuo, Y. & Ishizuka, M. 2006. Disambiguating personal names on the web using automatically extracted key phrases. In *Proceedings of the 17th European Conference on Artificial Intelligence*, 553–557. IOS Press.

Bollegala, D., Matsuo, Y. & Ishizuka, M. 2007. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*, 757–766. ACM.

Bu, F., Hao, Y. & Zhu, X. 2011. Semantic relationship discovery with Wikipedia structure. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence – Vol. 3 (IJCAI '11)*, Walsh, T. (ed.). AAAI Press, 1770–1775.

Budan, I. A. & Graeme, H. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics* **32**(1), 13–47.

Budanitsky, A. & Hirst, G. 2006. Evaluating Wordnet-based measures of lexical semantic relatedness. *Computational Linguistics* **32**(1), 13–47.

Chen, H. H., Lin, M. S. & Wei, Y. C. 2006. Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 1009–1016. Association for Computational Linguistics.

Chen, P., Ding, W., Bowes, C. & Brown, D. 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, 28–36. Association for Computational Linguistics.

Cilibrasi, R. L. & Vitanyi, P. 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* **19**(3), 370–383.

Duan, J. & Zeng, J. 2012. Computing semantic relatedness based on search result analysis. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology – Vol. 3*, 205–209. IEEE Computer Society.

Euzenat, J. & Shvaiko, P. 2013. *Ontology Matching*, 2nd edition. Springer-Verlag.

Feng, Y., Fani, H., Bagheri, E. & Jovanovic, J. 2015. Lexical semantic relatedness for Twitter analytics. In *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI 2015)*, 202–209. IEEE.

Ferrara, F. & Tasso, C. 2013. Evaluating the results of methods for computing semantic relatedness. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing – Part I (CICLing '13)*, 447–458. Springer-Verlag.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems* **20**(1), 116–131.

Gabrilovich, E. & Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, Sangal, R., Mehta, H. & Bagga, R. K. (eds). Morgan Kaufmann Publishers Inc., 1606–1611.

Gracia, J. & Mena, E. 2008. Web-based measure of semantic relatedness. In *Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE '08)*, Bailey, J., Maier, D., Schewe, K. D., Thalheim, B. & Wang, X. S. (eds). Springer-Verlag, 136–150.

Graham, M., Milanowski, A. & Miller, J. 2012. *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. http://files.eric.ed.gov/fulltext/ED532068.pdf.

Gruninger, M. & Kopena, J. B. 2005. Semantic integration through invariants. *AI Magazine* **26**(1), 11–20.

Gurevych, I. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP '05)*, Dale, R., Wong, K. F., Su, J. & Kwong, O. Y. (eds). Springer-Verlag, 767–778.

Gurevych, I. 2006. *Computing semantic relatedness across parts of speech*. Technical report, Department of Computer Science, Telecooperation, Darmstadt University of Technology.

Gurevych, I. & Niederlich, H. 2005. Computing semantic relatedness of GermaNet concepts. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Proceedings of the Workshop on Applications of GermaNet II at GLDV2005*, 462–474.

Hecht, B., Carton, S. H., Quaderi, M., Schöning, J., Raubal, M., Gergle, D. & Downey, D. 2012. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*, 415–424. ACM.

Hughes, T. & Ramage, D. 2007. Lexical semantic relatedness with random graph walks. In *Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, 581–589.

Jarmasz, M. & Szpakowicz, S. 2012a. Roget's thesaurus and semantic similarity. *arXiv preprint arXiv:1204.0245*.

Jarmasz, M. & Szpakowicz, S. 2012b. Roget's thesaurus: a lexical resource to treasure. *arXiv preprint arXiv:1204.0258*.

Jiang, J. J. & Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Karanastasi, A. & Christodoulakis, S. 2007. The OntoNL semantic relatedness measure for OWL ontologies. In *Proceedings of the 2nd International Conference on Digital Information Management, ICDIM '07*, 333–338. IEEE Computer Society.

Krizhanovsky, A. A. & Lin, F. 2009. Related terms search based on WordNet/Wiktionary and its application in ontology matching. *arXiv preprint arXiv:0907.2209*.

Leacock, C. & Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database* **49**(2), 265–283.

Leong, C. W. & Mihalcea, R. 2011. Measuring the semantic relatedness between words and images. In *Proceedings of the 9th International Conference on Computational Semantics*, 185–194. Association for Computational Linguistics.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86)*, DeBuys, V. (ed.). ACM, 24–26.

Li, Y., Bandar, Z. A. & McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* **15**(4), 871–882.

Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K. & Ishizuka, M. 2007. Polyphonet: an advanced social network extraction system. *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(4), 262–278.

Meyer, C. M. & Gurevych, I. 2012. To exhibit is not to loiter: a multilingual, sense-disambiguated Wiktionary for measuring verb similarity. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 1763–1780.

Mihalcea, R. & Moldovan, D. I. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*, 152–158. Association for Computational Linguistics.

Mika, P. 2007. Ontologies are us: a unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(1), 5–15.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS '13)*, 3111–3119. Curran Associates Inc.

Mikolov, T., Yih, W. T. & Zweig, G. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, 746–751. The Association for Computational Linguistics.

Milikic, N., Jovanovic, J. & Stankovic, M. 2011. Discovering the dynamics of terms' semantic relatedness through Twitter. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages*, 57–68.

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

Milne, D. 2007. Computing semantic relatedness using Wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.3604.

Mori, J., Ishizuka, M. & Matsuo, Y. 2007. Extracting keyphrases to represent relations in social networks from web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI '07)*, 2820–2825. Morgan Kaufmann Publishers Inc.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J. & Gómez-Rodríguez, A. 2015. Ontology matching. *Expert Systems With Applications* **42**(2), 949–971.

Patwardhan, S. & Pedersen, T. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense – Bringing Computational Linguistics and Psycholinguistics Together*, **1501**, 1–8.

Pedersen, T. 2012. Duluth: measuring degrees of relational similarity with the gloss vector measure of semantic relatedness. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval '12)*, 497–501. Association for Computational Linguistics.

Pedersen, T., Pakhomov, S. V., Patwardhan, S. & Chute, C. G. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* **40**(3), 288–299.

Pirró, G. 2012. REWOrD: semantic relatedness in the web of data. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI '12)*, 129–135. AAAI Press.

Polčicová, G. & Návrat, P. 2002. Semantic similarity in content-based filtering. In *Proceedings of the 6th East European Conference on Advances in Databases and Information Systems (ADBIS '02)*, Manolopoulos, Y. & Návrat, P. (eds). Springer-Verlag, 80–85.

Rada, R., Mili, H., Bicknell, E. & Blettner, M. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics* **19**(1), 17–30.

Radinsky, K., Agichtein, E., Gabrilovich, E. & Markovitch, S. 2011. A word at a time: computing word relatedness using Temporal Semantic Analysis. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, 337–346. ACM.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Vol. 1 (IJCAI '95)*, Mellish, C. S. (ed.). Morgan Kaufmann Publishers Inc., 448–453.

Resnik, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* **11**, 95–130.

Rubenstein, H. & Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM* **8**(10), 627–633.

Sabou, M., Gracia, J., Angeletou, S., d'Aquin, M. & Motta, E. 2007. Evaluating the semantic web: a task-based approach. In *Proceedings of the 6th International Semantic Web Conference, ISWC 2007*, 423–437. Springer-Verlag.

Sahami, M. & Heilman, T. D. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)*, 377–386. ACM.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics* **24**(1), 97–123.

Seco, N., Veale, T. & Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI '2004*, 1089–1090.

Spanakis, G., Siolas, G. & Stafylopatis, A. 2009. A hybrid web-based measure for computing semantic relatedness between words. In *Proceedings of the 2009 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI '09)*, 441–448. IEEE Computer Society.

Strube, M. & Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence – Vol. 2 (AAAI '06)*, Cohn, A. (ed.). AAAI Press, 1419–1424.

Taieb, M. A. H., Aouicha, M. B. & Hamadou, A. B. 2013. Computing semantic relatedness using Wikipedia features. *Knowledge-Based Systems* **50**, 260–278.

Turdakov, D. & Velikhov, P. 2008. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proceedings of the SYRCODIS 2008 Colloquium on Databases and Information Systems*. `http://ceur-ws.org/Vol-355/turdakov.pdf`.

Turney, P. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Committee on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, 313–320. Association for Computational Linguistics.

Turney, P. D. & Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**(1), 141–188.

Vélez, B., Weiss, R., Sheldon, M. A. & Gifford, D. K. 1997. Fast and effective query refinement. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, 6–15. ACM.

Wan, S. & Angryk, R. 2007. Measuring semantic similarity using Wordnet-based context vectors. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2007. ISIC*, 908–913. IEEE Computer Society.

Weng, J. & Lee, B. S. 2011. Event detection in Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, 401–408. Association for the Advancement of Artificial Intelligence.

Witten, I. & Milne, D. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, 25–30. AAAI Press.

Wu, H., Min, M. R. & Bai, B. 2014. Deep semantic embedding. In *Proceedings of Workshop on Semantic Matching in Information Retrieval Co-Located with the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 46–52.

Wu, Z. & Palmer, M. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL '94)*, 133–138. Association for Computational Linguistics.

Yang, D. & Powers, D. M. 2006. Verb similarity on the taxonomy of WordNet. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E. & Soroa, A. 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing*, 41–49. Association for Computational Linguistics.

Zarrinkalam, F., Fani, H., Bagheri, E. & Kahani, M. 2016. Inferring implicit topical interests on Twitter. In *Proceedings of the 38th European Conference on IR Research, ECIR 2016*, 479–491. Springer International Publishing.

Zesch, T. 2010. *Study of semantic relatedness of words using collaboratively constructed semantic resources*. PhD thesis, Technische Universität.

https://doi.org/10.1017/S0269888917000029

Zesch, T. & Gurevych, I. 2006. Automatically creating datasets for measures of semantic relatedness. In *Proceedings of the Workshop on Linguistic Distances (LD '06)*, 16–24. Association for Computational Linguistics.

Zesch, T. & Gurevych, I. 2010. The more the better? Assessing the influence of Wikipedia's growth on semantic relatedness measures. In *Proceedings of the Conference on Language Resources and Evaluation (LREC '10)*.

Zesch, T., Gurevych, I. & Mühlhäuser, M. 2007. Comparing Wikipedia and German Wordnet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies 2007: Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 205–208. Association for Computational Linguistics.

Zesch, T., Müller, C. & Gurevych, I. 2008. Using Wiktionary for computing semantic relatedness. In *Proceedings of the 23rd National Conference on Artificial Intelligence – Volume 2 (AAAI '08)*, Cohn, A. (ed.). AAAI Press, 861–866.

Zhao, Q., Hoi, S. C., Liu, T. Y., Bhowmick, S. S., Lyu, M. R. & Ma, W. Y. 2006. Time-dependent semantic similarity measure of queries using historical click-through data. In *Proceedings of the 15th International Conference on World Wide Web*, 543–552. ACM.

Zhou, W., Wang, H., Chao, J., Zhang, W. & Yu, Y. 2012. LODDO: using linked open data description overlap to measure semantic relatedness between named entities. In *Proceedings of the 2011 Joint International Conference on The Semantic Web (JIST '11)*, Pan, J. Z., Chen, H., Kim, H. G., Li, J. & Wu, Z. (eds). Springer-Verlag, 268–283.

https://doi.org/10.1017/S0269888917000029