Natural Language Engineering **19** (4): 411–479. © Cambridge University Press 2012
doi:10.1017/S1351324912000125

Recent advances in methods of lexical semantic relatedness – a survey

ZIQI ZHANG, ANNA LISA GENTILE and FABIO CIRAVEGNA

Department of Computer Science, University of Sheffield
211 Portobello, Regent Court, Sheffield, UK, S1 4DP
e-mail: {z.zhang,a.l.gentile,f.ciravegna}@dcs.shef.ac.uk

(*Received 7 April 2011; revised 16 December 2011; accepted 16 March 2012; first published online 4 May 2012*)

Abstract

Measuring lexical semantic relatedness is an important task in Natural Language Processing (NLP). It is often a prerequisite to many complex NLP tasks. Despite an extensive amount of work dedicated to this area of research, there is a lack of an up-to-date survey in the field. This paper aims to address this issue with a study that is focused on four perspectives: (i) a comparative analysis of background information resources that are essential for measuring lexical semantic relatedness; (ii) a review of the literature with a focus on recent methods that are not covered in previous surveys; (iii) discussion of the studies in the biomedical domain where novel methods have been introduced but inadequately communicated across the domain boundaries; and (iv) an evaluation of lexical semantic relatedness methods and a discussion of useful lessons for the development and application of such methods. In addition, we discuss a number of issues in this field and suggest future research directions. It is believed that this work will be a valuable reference to researchers of lexical semantic relatedness and substantially support the research activities in this field.

1 Introduction

Measuring semantic relatedness between lexically expressed concepts is an important task in Natural Language Processing (NLP). It is often a pre-processing step to many NLP applications such as Word Sense Disambiguation (Leacock and Chodorow 1998; Han and Zhao 2010), Named Entity Recognition (Kliegr *et al.* 2008), clustering (Matsuo *et al.* 2006; Bollegala, Matsuo and Ishizuka 2007), and Information Retrieval (Finkelstein *et al.* 2002). In the biomedical domain, it is an important technique for discovering functional relations between genes, and constructing lexical resources (Ye *et al.* 2005; Al-Mubaid and Nguyen 2006; Wu *et al.* 2006; Li *et al.* 2010). Extensive research has been carried out to study the methods of lexical semantic relatedness. These methods exploit lexical and semantic information of words and concepts usually encoded in some background information resources.

An up-to-date review of the literature is currently unavailable. Although some have attempted to address this issue (Budanitsky and Hirst 2006; Pesquita *et al.*

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

2009; Zesch and Gurevych 2010a), the scope of these studies is limited in a number of ways. First, they focus on a subset of methods, such as those based on a specific background information resource about concepts and words. A summary of work across such boundaries is lacking. Second, collaborative resources such as Wikipedia and Wiktionary have been studied substantially in this field in recent years and a large number of methods (Strube and Ponzetto 2006; Gabrilovich and Markovitch 2007; Milne and Witten 2008; Turdakov and Velikhov 2008; Zesch, Müller and Gurevych 2008b; Hassan and Mihalcea 2009; Yazdani and Popescu-Belis 2010; Radinsky *et al.* 2011; Zhang, Gentile and Ciravegna 2011) have been introduced to exploit these resources. A summary of these studies is currently lacking. Third, we identify a gap between the research of lexical semantic relatedness in general and specific domains, particularly the biomedical domain. On the one hand, an increasing number of new methods have been proposed for general purposes; however, only a very small amount of these have been adapted to the biomedical domain. On the other hand, in the biomedical community, significant effort has been spent on developing novel methods tailored to specific biomedical resources and datasets. Despite the differences in terms of domains and resources used, we believe that both communities can benefit significantly by sharing the knowledge and lessons learnt. Unfortunately, work from the two communities is rarely compared or transferred. As a result of this, it has been noticed that very similar methods have been introduced in different contexts, potentially costing expensive research effort.

This work surveys the literature on methods for measuring lexical semantic relatedness, focusing on recent studies that have not been covered in the previous surveys, and connecting work from both general and biomedical domains. We discuss the rationales of the methods, analyze their connections, and present evaluations based on previously published resources. This survey aims to draw lessons on the development, evaluation, and application of lexical semantic relatedness methods, as well as to identify remaining challenges in the research field. The remainder of this survey is structured as follows: Section 2 introduces key notions and terminologies used in this work. Section 3 discusses background information resources that are used for measuring lexical semantic relatedness. Section 4 discusses methods for lexical semantic relatedness, focusing on recent proposals and those from the biomedical domain. Section 5 describes evaluation methods and standards, and summarizes evaluation results based on published resources. Section 6 draws lessons from the discussions. Section 7 briefly outlines reusable tools for the research into, and application of, semantic relatedness. Section 8 concludes this work and discusses future research directions.

2 Terminology and definitions

In the literature on lexical semantic relatedness, the term *semantic relatedness* is often confused with three different but relevant terms: *semantic similarity*, *semantic distance*, and *distributional similarity*. Semantic relatedness essentially describes the strength of semantic association between two concepts, or their lexical realizations. It encompasses a variety of relations between concepts, including the classical relations

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

such as hypernymy, hyponymy, meronymy, antonymy, synonymy, and any other 'nonclassical relations' (Morris and Hirst 2004) and 'implicit connections' (Zesch and Gurevych 2010a). *Semantic similarity* is a specific case of relatedness, where the sense of relatedness is dependent on the 'degree of synonymy' (Weeds 2003), which is usually accounted by classical relations. Concepts that are semantically related are not necessarily similar, such as *car* and *fuel*. Another example is that antonyms are considered to be semantically related, such as *beautiful* and *ugly*; however, they are dissimilar. Computational applications typically require relatedness rather than similarity (Budanitsky and Hirst 2006). The term *semantic distance* has been used in the literature to refer to the inverse of semantic relatedness or similarity. Concepts that are semantically similar or related are considered to be semantically close to each other, thus denoting a sense of distance. It is also worth noting that although it is generally agreed that semantic relatedness is symmetric, this is not always true for semantic similarity (Tversky 1977). Asymmetric similarity is often perceived between a concept and its superclass concepts. Similarity from a concept to its superclass is usually considered greater than the opposite. For example, a *pear* is similar to a *fruit* is more agreeable than a *fruit* is similar to a *pear*. However, the literature has predominantly taken the assumption of symmetric semantic relatedness and similarity.

Theoretically, semantic relatedness is defined for concepts. However, in practice, concepts are lexicalized as *words* or *phrases*. Due to the polysemy of language, a word or phrase (e.g., *bank*) may have multiple senses, each of which refers to a different meaning (*bank, the financial institution,* or *bank, the land adjoining a body of water*), or *concept* (Budanitsky and Hirst 2006). Some literature (Zesch and Gurevych 2010a) has used 'term' to refer to both words and phrases; while in the biomedical domain a 'term' sometimes refers to a nonambiguous sequence of tokens denoting a unique specialized concept. In this survey, when we say *term*, we adopt the same notion as that in Zesch and Gurevych (2010a), namely a non-sense-tagged word or sequence of words, which may have multiple meanings corresponding to different concepts. Meanwhile, to be consistent with the literature, we may occasionally use 'term' to refer to a unique concept in the biomedical domain. When we do so we will clarify this.

Following these notions, semantic relatedness between two terms is typically evaluated based on their underlying concepts. In addition, methods for measuring *distributional similarity* of words (Weeds 2003) have been largely applied as a proxy to address lexical semantic relatedness. In the literature, they are sometimes used interchangeably with *word similarity*, or *co-occurrence similarity*. Briefly, distributional similarity between two words is based on the extent to which the two words tend to occur in similar contexts. By this definition, distributional similarity does not strictly adhere to the notion of lexical semantic relatedness, and its application to assessing semantic relatedness between words has been controversial (Weeds 2003; Budanitsky and Hirst 2006). Despite these debates, substantial work has been carried out to develop and apply distributional similarity methods to address the issue of semantic relatedness and related tasks, which has proved effective. For this reason, this survey also covers studies that fall under this category but are tailored for the task of lexical semantic relatedness. For a comprehensive discussion of distributional similarity, readers are advised to refer to Weeds (2003).

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

3 Background information resources

Measuring lexical semantic relatedness generally requires certain *background information* about the concepts or terms. Such information is often encoded in structured and semi-structured *knowledge bases* that form a graph of concepts, which are lexicalized and indexed by their word forms. Concepts are interconnected by *links* or *edges* that denote a certain sense of semantic relations. Depending on the semantics of the links and the shape of the graph, we will use the following: the term *taxonomy* to refer to a hierarchical structure, in which nodes are organized by the *generalization–specialization* relationship; the term *ontology* to refer to a taxonomic structure enriched with other semantic relationships such as antonymy and synonymy, and class properties or attributes; and the term *semantic graph* or *semantic network* to refer to any kind of concept graphs connected by any semantic or loose associative relations. Following these notions, taxonomy is a specialization of ontology, and they both are specific types of semantic graph or network. Semantic graphs may also contain multiple taxonomies or ontologies.

In an analogy, distributional similarity can be considered to employ background information of words in the form of their contexts, which are derived from a large corpus. We will refer to this source of background information as *unstructured corpora* in the sense that the documents do not provide sense-tagging of words, or explicitly organize words or concepts in a structured way encoding their associations as it is in knowledge bases.

The remainder of this section discusses knowledge bases and unstructured corpora most frequently used in the task of measuring lexical semantic relatedness. We explain what and how information can be extracted from these resources and used for this task.

3.1 Knowledge bases

Examples of knowledge bases include dictionaries, thesauri, wordnets, and encyclopedic resources. Some methods, especially earlier ones, have employed dictionaries and thesauri, such as Morris and Hirst (1991), Kozima and Furugori (1993), and Jarmasz and Szpakowicz (2003). Most methods employ wordnets, and encyclopedic resources. Among these, the most frequently used general purpose knowledge bases include WordNet (Fellbaum 1998), Wiktionary,[1] and Wikipedia.[2] In the biomedical domain, domain-specific knowledge bases are also available, such as the Gene Ontology (GO) (The Gene Ontology Consortium 2005) and the Medical Subject Headings[3] (MeSH) biomedical vocabulary resource.

3.1.1 General purpose knowledge bases

WordNet is a lexicalized ontology of English words. It groups nouns, verbs, adjectives, and adverbs into *synsets*, each expressing a distinct concept. Searching for a word in

[1] http://www.wiktionary.org/, last retrieved on 16 March 2012.
[2] http://www.wikipedia.org/, last retrieved on 16 March 2012.
[3] http://www.nlm.nih.gov/mesh/, last retrieved on 16 March 2012.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

WordNet may return multiple synsets corresponding to different senses or concepts. WordNet provides polysemy counts of a word by encoding the frequency of the word being found annotated with each synset in a sense-tagged sample corpus. Each concept in WordNet is provided with a short definition called *gloss*, and is connected to other concepts by a set of semantic relations depending on the word class, such as hypernymy and meronymy for nouns, hypernymy and entailment for verbs, and synonymy and antonymy for adjectives.

Since WordNet is designed to provide complete coverage of common, open-class English words, it has little or no coverage of vocabularies from specialized domains, and very limited coverage of proper nouns. This may hinder its application to domain-specific contexts and tasks required to deal with proper nouns (Hirst and St-Onge 1998; Strube and Ponzetto 2006). WordNet equivalents in other languages have also been created. For example, the German equivalent GermaNet (Kunze and Lemnitzer 2002) has also been used in measuring semantic relatedness for German words (Zesch *et al.* 2008b; Zesch and Gurevych 2010a).

WordNet has been one of the most popular background knowledge bases for the studies of semantic relatedness. The majority of WordNet-based methods make use of the taxonomic structure (Wu and Palmer 1994; Leacock and Chodorow 1998; Li, Bandar and McLean 2003), while some exploit the entire semantic graph taking into account all kinds of relations (Hughes and Ramage 2007; Agirre *et al.* 2009). WordNet gloss and semantic relations have also been exploited as features in building concept vectors (Zesch and Gurevych 2010a; Zhang *et al.* 2011).

Wiktionary is a multilingual free dictionary built and maintained by collaborative effort. As of January 2012, it covers over 150 languages and more than 2.8 million entries. Wiktionary shows many commonalities with WordNet: each entry in Wiktionary is an article page about a term and distinguishes one or more word classes. Each word class has one or more senses that correspond to concepts. Each concept is provided with a short definition (similar to WordNet gloss) often accompanied by example sentences. Wiktionary also defines lexical semantic relations that are available in WordNet, such as hypernymy, hyponymy, coordinate terms, synonymy, and antonymy. However, they are encoded at the level of word class rather than concepts. In addition to these commonalities shared with WordNet, it also encodes information such as alternative forms and etymology at the level of terms; and derived terms and translation at the level of word class.

Meyer and Gurevych (2010) compared resource coverage of English Wiktionary against WordNet. They showed that in general Wiktionary encodes twice the amount of words than WordNet. Wiktionary outnumbers WordNet by covering a broader range of word classes, a larger number of abbreviations, numerals, symbols, and proper nouns. Wiktionary covers a large number of word inflectional forms (nearly 30% of Wiktionary) and neologisms, which are unavailable in WordNet. However, about half the amount of WordNet lexicons are missing in the English Wiktionary, of which 50% are found to be Latin words belonging to scientific domains. Meyer and Gurevych also studied the word sense distribution over different word classes and sense alignment in the two resources. They concluded that the distribution in both resources is very similar, despite that on average WordNet encodes more word senses

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

for verbs whereas Wiktionary encodes more word senses for nouns. Wiktionary has better coverage of slang-related and domain-specific senses, as well as word senses for rarely used terms. Using a sample corpus, they discovered that both knowledge bases share many word senses for words with a medium language frequency; while Wiktionary encodes a large number of word senses for words with a high frequency.

Despite such arguably favorable observations for Wiktionary, Navarro *et al.* (2009) show that Wiktionary suffers from issues such as uneven density of knowledge, imbalanced coverage of different languages, and a sparse synonym network. For example, the lexical–semantic information is not always encoded for any words belonging to the same word class. And the amount of encoded information is largely imbalanced. Such issues are believed to be caused by the nature of collaborative authoring, where the creation of knowledge is largely driven by user interests.

Compared with WordNet, the use of Wiktionary in semantic relatedness is rarely investigated, although some pilot work (Zesch *et al.* 2008b) has been made to adapt several WordNet-based methods to the Wiktionary semantic graph.

Wikipedia is a multilingual encyclopedia created and maintained by collaborative effort. Article pages in Wikipedia describe a vast amount of proper nouns (or entities) and concepts. They do not have a dedicated section of definitions similar to WordNet gloss or Wiktionary definitions. However, research has assumed that the first paragraph of a Wikipedia article provides definitional details and the first sentence often gives a short definition. Wikipedia articles are hyperlinked; however, the links are not typed. They denote rather general semantic associations and thus create a loosely connected semantic graph. Articles are tagged with multiple category labels, which are general concepts organized in a hierarchical structure, creating a category tree generally resembling the broader and narrower sense of relation between categories. In addition, Wikipedia groups synonyms and aliases using the mechanism of 'redirect' – an article page may be linked to a number of alternative names denoting a sense of synonyms, which when searched, will always be redirected to the uniform article page. Polysemous names and phrases are encoded in separate 'disambiguation' pages, which list different meanings with links to corresponding article pages. For many Wikipedia pages, a tabular 'infobox' of additional metadata is available, usually presenting fact-like information of certain types that are common to similar articles.

Wikipedia offers several advantages over WordNet and Wiktionary. Most of all, it covers a substantial amount of proper nouns and concepts, as well as domain-specific vocabularies. Holloway, Bozicevic and Börner (2007) showed that by 2005, Wikipedia already contained 1,069 disconnected clusters of categories of articles each denoting a distinctive subject. Milne, Medelyan and Witten (2006) showed that in the domain of food and agriculture, Wikipedia provides excellent coverage of domain terminology and semantic relations that rivals a professional thesaurus. Halavais and Lackaff (2008) in an analysis of topical coverage of Wikipedia by comparing printed books against Wikipedia articles concluded that the coverage of topic-specific knowledge is generally good in Wikipedia. In addition, the denser connections between article pages and categories, as well as longer content also imply richer lexical semantic information.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

However, Wikipedia does not annotate article hyperlinks by semantic relations. Likewise, the hierarchical structure of category tree rather represents a loose folksonomy than a strict taxonomy (Strube and Ponzetto 2006; Ponzetto and Strube 2011), since it contains hypernymy relation as well as other relations such as meronymy. Such relations are not explicitly distinguished. Also, because of the encyclopedic purpose of Wikipedia rather than a lexical knowledge base, its content may be biased toward specialized concepts and instances rather than lexicographic senses of words. In particular, verbs are largely under-represented, as shown in Zesch and Gurevych (2010a). For example, the closest entry matching the word *win* and its verb sense is the article on *victory*, while all the other articles describe domain-specific concepts or entities referred to by the same word.

Wikipedia has been a popular choice of knowledge base in recent work. Similar to Wiktionary, many studied the adaptation of WordNet-based methods to the hierarchical category graph and article contents (Strube and Ponzetto 2006; Zesch and Gurevych 2010a). The link structure of Wikipedia articles has also been exploited by a number of graph-based approaches (Yeh *et al.* 2009; Yazdani and Popescu-Belis 2010).

3.1.2 Biomedical knowledge bases

Despite the extensively studied general-purpose knowledge bases discussed above, studies in the biomedical domain generally prefer domain-specific knowledge bases due to their comprehensive coverage of biomedical knowledge. Such resources are exceptionally abundant in this domain, thanks to decades of extensive research devoted to this area. However, domain-specific knowledge bases in other technical domains are generally scarce. Biomedical knowledge bases are usually a structured vocabulary of technical terms, which often denote unique, specialized concepts. Several frequently used biomedical knowledge bases are briefly introduced below.

The *Gene Ontology* is the most often used knowledge base for semantic relatedness in the biomedical domain. GO provides 'an ontology of defined terms representing gene product properties' (GeneOntology.org), where each term refers to a unique concept. The terms are further divided into three sub-domains: cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding and catalytic activities; and biological process, operations, or sets of molecular events relevant to the functioning of cells, tissues, organs, and organisms. For example, the term *heart contraction* is a concept of a biological process representing a property of the gene product *Actin, alpha cardiac muscle 1*. Each term is assigned a unique identifier; a term name which is a word or word sequence; a short definition similar to that of WordNet gloss; and a name space label indicating the sub-domain that the term belongs to. Terms are interlinked with other terms within or across sub-domains, by relations such as synonymy, hypernymy, meronymy, and domain-specific relations such as regulation.

The majority of studies in the biomedical domain exploit the GO as the background information resource (Pesquita *et al.* 2009). Also, many have exploited the

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

semantic similarity between GO terms to assess similarity between genes or gene products annotated by these terms. The idea is that genes are similar if they share same properties, which are described by GO terms. Due to the similarity in its structure with WordNet, many of these methods have adapted WordNet-based methods or are developed based on similar rationales.

The *MeSH* is a comprehensive controlled vocabulary resource for indexing articles and books in the biomedical domain. It is structured as a hierarchy of 'descriptors', each of which is essentially a subject heading designed for the purpose of indexing. A descriptor is accompanied with a definition of the description; a list of synonyms or very similar names known as 'entry terms', in the sense that the same descriptor can be looked up using such terms. This is similar to the Wikipedia redirect mechanism. The descriptors are split into sixteen categories representing sub-topics such as organisms, diseases, and drugs. Descriptors of each category are organized as sub-hierarchies from most general to most specific levels. A MeSH descriptor may appear in multiple places in a hierarchy.

The *Systematized Nomenclature of Medicine – Clinical Terms*[4] *(SNOMED-CT)* is a controlled vocabulary of medical terminology. The terminology is organized into thirteen hierarchies based on topics such as clinical findings, procedures, organisms, and substances, containing over 1 million concepts. Each concept is assigned a unique identifier, and provided with multiple 'descriptions', each of which is a name used to refer the concept. Thus, these are similar to the 'entry terms' in MeSH. The descriptions are divided into three types: a unique 'Fully Specified Name (FSN) a "preferred term', and one or multiple synonyms. Preferred terms and synonyms are names that are not unique to the concept, but can be shared by multiple concepts. Concepts in SNOMED-CT are organized as taxonomies following the hypernym relation. In addition, a number of domain-specific relations (e.g., 'due to', 'causative agent') are defined to connect concepts.

The *Unified Medical Language System*[5] *(UMLS)* is a resource that maps a wide range of biomedical knowledge bases. Initiated in 1986, the objective is to create a structured knowledge base of biomedical science to facilitate the development of computer systems that can understand biomedical language. It contains three main knowledge bases: Metathesaurus, Semantic Network, and SPECIALIST lexicon. The Metathesaurus contains over 1 million biomedical concepts and 5 million concept names integrated from over one hundred knowledge bases (source vocabularies), including GO, MeSH, and SNOMED-CT. One of its main purposes is to group different names for the same concept from different source vocabularies. Each concept is assigned a unique identifier, one or multiple concept names, and pointers to their source vocabularies. Many relationships are encoded between concepts, including the hypernymy, meronymy, and synonymy relations. Relations encoded in source vocabularies are also retained. The Semantic Network defines a set of subject categories called semantic types, such as organisms, biologic functions, and

[4] http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/snomed, last retrieved on 16 March 2012.
[5] http://www.nlm.nih.gov/research/umls/, last retrieved on 16 March 2012.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

chemicals. They are organized into a hierarchy representing the hypernymy relation, and also interlinked by many other non-hierarchical relations. The SPECIALIST lexicon is intended to be an English lexicon of both common English words and biomedical vocabulary. An entry is defined for each word or phrase, and records the syntactic, morphological, and orthographical information that can be used by NLP systems.

The Unified Medical Language System has been used in a wide range of NLP tasks in the biomedical domain. However, its usage in semantic relatedness is limited to the scope of individual knowledge bases it has integrated. The usefulness of the Semantic Network, for example, has not been investigated for this task. McInnes, Pedersen and Pakhomov (2009) developed a toolkit that allows adapting a number of WordNet-based methods to the UMLS Metathesaurus only. The Semantic Network is not used.

The *Human Phenotype Ontology* (HPO) is a vocabulary of approximately 9,000 terms referring to phenotypic abnormalities encountered in human disease (Kohler *et al.* 2009). A term in HPO denotes a unique concept, which has a unique identifier and a unique name. Some terms are provided with a short definition, and many terms have one or multiple cross-reference identifiers to UMLS concepts. Terms in HPO are grouped hierarchically following the hypernymy relation, where multiple parents of a term are allowed. Synonymy and meronymy relations are also defined between terms.

The *Chemical Entities of Biological Interest (ChEBI)* is 'an ontology of molecular entities focused on small chemical compounds' (Degtyarenko *et al.* 2007). It contains four sub-ontologies, each addressing one of the four topics: molecular structure, which classifies molecular entities or their parts according to their structure; biological role, which classifies entities based on their role in a biological context, such as antibiotics; application, which classifies entities or their parts by their intended use by humans, such as drugs; and subatomic particle that classifies particles smaller than atoms. Each entity is assigned a unique identifier and a ChEBI name, which is sometimes ambiguous and can refer to multiple entities. Entities are primarily organized following the hypernym relation. Synonymy and meronymy relations between entities are also encoded.

Table 1 summarizes the knowledge bases discussed above.

3.2 Unstructured corpora

As mentioned before, unstructured corpora can be considered the background information resource for distributional similarity methods. Some semantic relatedness methods also employ unstructured corpora in certain ways. Unlike knowledge bases, unstructured corpora do not provide sense tagging of words or define lexical semantic relations in an explicit way. As a result, background information is collected at the level of words rather than concepts, and connects words in a rather implicit way. With distributional similarity methods, this is often in the form of textual contexts of words or their co-occurring behaviors within certain contexts based on a sufficiently

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Table 1. *Summary of general purpose and domain-specific knowledge bases*

Knowledge bases	Coverage of knowledge		Structure and content
	Focus of coverage	Limited coverage	
WordNet (GN)	Common English words including nouns, verbs, adjectives and adverbs.	Limited coverage of specialized vocabularies, very few proper nouns.	• Gloss: a definitional description. • Synsets are linked by semantic and lexical relations; such as hypernymy, meronymy, synonymy, coordinate etc.
Wiktionary (GN)	Nouns, verbs, adjectives, adverbs, other word classes; abbreviations, neologisms, inflectional words.	Covers some specialized vocabularies and more proper nouns than WordNet.	• Definitional description with example sentences. • Most WordNet lexical and semantic relations are encoded, plus additional information such as etymology.
Wikipedia (GN)	Proper nouns and concepts, covering a broad range of topics.	Not focusing on lexicographic senses of words, which can be under-represented (e.g., verbs).	• Dedicated content pages for concepts and proper nouns. • Heavily hyperlinked article pages. • Hierarchical category tree for classification of articles. • Redirect system grouping synonyms and aliases.
GO (DS)	Concepts and vocabularies representing gene product properties.		• Definitional description. • Hypernymy, meronymy, synonymy, and domain specific relations.
MeSH (DS)	Descriptors denoting unique concepts of the medical domain.		• Definitional description and synonymous terms, which are names used to refer to the same descriptor concept. • Hierarchical structure representing generalization/specification.
SNOMED-CT (DS)	Concepts and vocabularies for the clinical domain.		• Synonymous terms (similar to those in MeSH). • Concepts linked by hypernymy and domain specific relations.
ULMS (DS)	Concepts and resources vocabularies integrated from other biomedical resources.		• Metathesaurus maps concepts across over 100 knowledge bases. • Semantic Network defines a category tree to classify concepts. • SPECIALIST lexicon defines syntactic, morphological and orthographic information of words.
HPO (DS)	Concepts and vocabulary of phenotypic abnormalities.		• Definitional descriptions, hypernymy, meornymy and synonymy relations. • Cross-reference to UMLS concepts.
ChEBI (DS)	Molecular entities focused on small chemical compounds.		• Entities linked by hypernymy, meronymy and synonymy relations.

GN: general purpose; DS: domain specific purpose.

large collection of documents. The hypothesis is that words that tend to occur in similar contexts are similar.

Some recent approaches have proposed mining lexical semantic network of words from unstructured document collections such that methods that use structured knowledge bases can be applied. For example, Harrington (2010) and Wojtinnek and Pulman (2011) proposed to parse a corpus to build connected graphs of words based on their syntactic relations, and exploit the link structure using graph-based algorithms to measure semantic relatedness between words. Pozo, Pazos and Valencia (2008) built a hierarchy of words by applying hierarchical clustering algorithms using the corpus, and applied methods that are based on hierarchical structures. Details of these will be discussed in Section 4.

3.2.1 General-purpose corpora

A large number of general-purpose document collections have been compiled for use in NLP research and applications. The most often used include the Brown corpus (Kucera and Francis 1967), the British National Corpus (The BNC Consortium 2007), the Penn Treebank corpus (Marcus, Marcinkiewicz and Santorini 1993), the Reuters corpus (Rose, Stevenson and Whitehead 2002), and the newswire articles published by Associated Press (AP newswire articles) and Wall Street Journals (WSJ), some of which are archived by Harman and Liberman (1993). Each of these compiles different document resources of various topics to the order of millions of words. They have been used to build statistical models of word contexts and co-occurrence behaviors used by distributional similarity methods such as Lee (1999) and Curran and Moens (2002). In addition, these are also used for modelling word usage patterns of English language in the general domain in methods such as Resnik (1995) and Lin (1998b).

Recently, an increasing number of approaches (Chen, Lin and Wei 2006; Matsuo *et al.* 2006; Cilibrasi and Vitanyi 2007; Gracia and Mena 2008) have explored the Web as the source of unstructured documents for distributional similarity methods. Typically, queries are composed based on the words in question and are used to retrieve documents that are likely to contain co-occurrences of words. Compared to precompiled document collections, such approaches can benefit from the sheer size of the Web, which generally provides a better coverage. However, they may also suffer from limitations inherited from search engines, such as limited query syntax, and potentially misused counting that is intended for number of pages rather than instances (Kilgarriff 2007).

3.2.2 Biomedical corpora

In the biomedical domain, corpora are usually pre-processed with distributional statistics stored in relational databases. The statistics are gathered for domain-specific vocabularies, usually terms defined in biomedical knowledge bases such as the GO. Such databases include the SWISS-PROT/UniProtKB or SP/UP (Boutet *et al.* 2007), the Saccharomyces Genome Database (SGD) (Cherry *et al.* 1998), the

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

InterPro protein sequence analysis and classification (Hunter *et al.* 2009), the Gene Ontology Annotation (GOA) database (Camon *et al.* 2004), the FlyBase database (McQuilton *et al.* 2011), the Mayo Clinic Corpus of Clinic Notes (Pakhomov, Coden and Chute 2004), the Online Mendelian Inheritance in Man (OMIM) database (McKusick 1998), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (Kanehisa and Goto 2006). Unlike counting word frequencies in a general corpus, for many of these databases (e.g., SP/UP, SGD, and GOA) the statistics are not based on the lexical realization of concepts, but according to their usage in *gene product annotations*. Typically, biomedical terms are used to annotate mentions of gene products in the corpus if they represent a property of gene products. Thus, the frequency of a term is determined by the frequencies of the gene products it annotates, and terms are said to co-occur if they are used to annotate same gene products.

In addition, large amounts of biomedical texts are published regularly by PubMed.[6] However, the majority of methods that use corpus statistics in the biomedical domain exploit one or multiple databases introduced above rather than document collections.

4 Methods of lexical semantic relatedness

This section discusses methods for measuring lexical semantic relatedness. The scope of the discussion will include studies from both the general and the biomedical domains, which will be presented from a general point of view. The focus of the discussion will be the rationales and connections of different methods, particularly for recent studies and work from the biomedical domain, where new methods have been introduced but have not been covered by previous surveys. Classic methods that have been thoroughly reviewed in previous work, their simple derivatives, and adaptations to other languages or resources are briefly described with pointers to related work.

We will use the term 'general purpose' or 'generic', or 'general domain' methods to refer to methods that were *originally* developed using general-purpose background information resources; and the term 'domain-specific' methods to refer to methods that were *originally* developed using domain-specific background information resources. As we shall see, both generic and domain-specific methods share the same or similar rationales, and use background information resources in similar ways. As a result, adaptation of generic methods to domain-specific background information resources should be generally straightforward and domain-specific methods can also be generalized for general purpose.

We will cover methods of semantic relatedness, as well as distributional similarity methods that have been used as a proxy for this purpose. In formulae, we use **SemRel** to denote semantic relatedness, and **DistSim** to denote distributional similarity. Many methods specifically address semantic similarity whereas some measure semantic distance as the inverse of relatedness or similarity. In the first case they are denoted as **SemSim**; in the latter case they are denoted as **SemDist**. Also, some methods

[6] http://www.ncbi.nlm.nih.gov/pubmed/, last retrieved on 16 March 2012.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

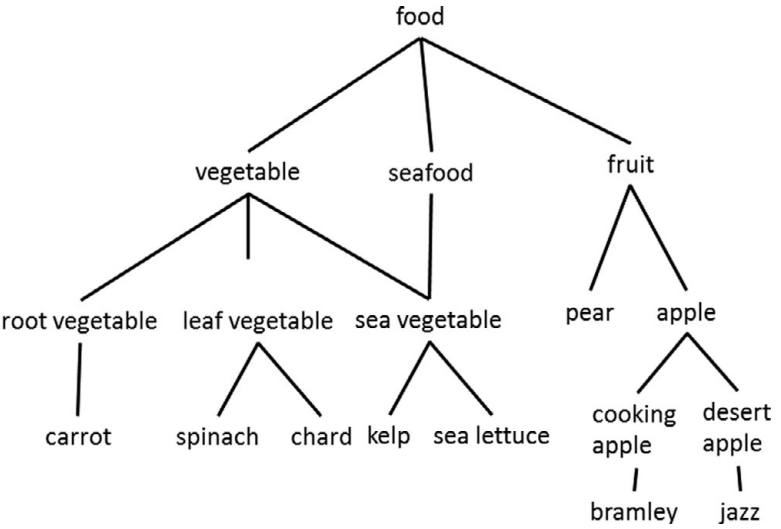


Fig. 1. An arbitrary example taxonomy used in this survey.

apply to concepts, while others apply to words or phrases. We use c to denote a concept, w to denote a polysemous word, phrase, or term. Concepts that are represented by the same word are denoted as $c \in C(w)$. Thus, $SemRel(c_1, c_2)$ denotes semantic relatedness between concepts c_1 and c_2 whereas $SemRel(w_1, w_2)$ denotes relatedness between words w_1 and w_2.

Given a pair of terms that can be polysemous, semantic relatedness methods typically calculate relatedness between their underlying concepts (i.e., $C(w)$), and then adopt a simple strategy to derive an aggregated score for their lexical expressions. The literature largely differs in terms of the approaches to measuring concept relatedness, while the method for deriving term relatedness is generally based on some rather *de facto* practice. Three techniques are commonly used for this purpose: (1) maximum pairwise concept relatedness, which assigns the maximum relatedness score obtained for every pair of $c_1 \in C(w_1)$ and $c_2 \in C(w_2)$; (2) average pairwise concept relatedness, which takes the average of relatedness scores obtained for every concept pair; and (3) sum of pairwise concept relatedness. To avoid repetition, in the remainder of this section, we will not explain for each individual method its choice of method for deriving word relatedness from concept relatedness.

Semantic relatedness – particularly similarity – is often measured with respect to a taxonomic structure, denoted by T. Figure 1 shows an example taxonomy of arbitrary *food* concepts. A number of common notions shared by these methods are defined below.

1. node – a node corresponds to a concept in the semantic graph. We will use node and concept interchangeably when describing taxonomies and semantic graphs.
2. edge – a single link connecting two adjacent nodes in a semantic graph. Where a semantic graph connects nodes by certain relations, we say that the type of an edge is defined by the relation it represents. For example, the edge between

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

vegetable and *food* in the example taxonomy represents the *IS-A* relation between the two nodes.

3. root – the root node in T, in this case, the node *food*.
4. *parent*(c) – a function that returns the parent node(s) of a node. For example, *parent*(*chard*) = {*leaf vegetable*}, and *parent*(*sea vegetable*) = {*vegetable, seafood*}. Some literature has used 'parent' to refer to any node that subsumes a node in a taxonomic structure, in which sense *parent*(c) returns the collection of nodes by following the taxonomic link from c to *root*. In this work, we use *subsumer*(c) to represent this case. Thus, when we say 'parent(s) of a concept' we always refer to the concept(s) that immediately subsumes the concept in the taxonomy.
5. *subsumer*(c) – a function that returns the subsumers of a node in a recursive manner. For example *subsumer*(*pear*)= {*fruit, food*}, and *subsumer*(*kelp*) = {*vegetable, seafood, food, sea vegetable*}. 'Subsumer' is also used interchangeably with 'ancestor', or 'hypernym' in the following discussion.
6. *child*(c) – a function that returns the child node(s) of a node. For example, *child*(*food*) = {*vegetable, fruit, seafood*}. Similar to the definition of *parent*(c), in this work, when we say 'child of a concept' we always refer to the concept(s) that are immediately subsumed by the concept in the taxonomy.
7. *descendant*(c) – a function that returns the child nodes of a node in a recursive manner. For example, *descendant*(*leaf vegetable*) = {*spinach, chard*}, and *descendant*(*vegetable*) = {*leaf vegetable, spinach, chard, sea vegetable, kelp, sea lettuce, root vegetable, carrot*}.
8. $cs(c_1, c_2)$ – returns the shared or common subsumers of concepts c_1, c_2. In mathematical terms, $cs(c_1, c_2) = subsumer(c_1) \bigcap subsumer(c_2)$. In some literature, this is called *common ancestor*.
9. $lcs(c_1, c_2)$ – returns the *least common subsumers* (*lcs*) of concepts c_1, c_2. There are different definitions of *lcs* in the literature. The majority of these adopt a definition by Resnik (1995), which defines *lcs* as the member in $cs(c_1, c_2)$ at the lowest level of taxonomy. Thus, $lcs(chard, kelp)$ = {*vegetable*}, and *lcs*(*kelp, sea lettuce*) = {*sea vegetable*}. In some literature this is called *most specific subsumer* (Budanitsky and Hirst 2006), or *lowest common ancestor* (Schickel-Zuber and Faltings 2007). Following this definition, theoretically two concepts may have multiple *lcs*, which could happen if each concept has multiple parents and two of these are shared between them. In practice, this is not very common. For example, a concept in WordNet has on average 1.03 parents according to Schickel-Zuber and Faltings (2007). Despite this low possibility, the authors proposed a revised definition to resolve such cases, and their work is the only one that applies a different definition of *lcs*. This will be introduced in Section 4.1.

Additional notations specific to each method will be introduced as they are encountered.

To facilitate the discussion, we classify semantic relatedness methods into several categories based on their fundamental rationales following the scheme introduced in

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Zesch and Gurevych (2010a). The authors classified existing methods into *path-based* (Section 4.1), *Information Content (IC)-based* (Section 4.2), *gloss-based* (Section 4.3), and *vector-based* (Section 4.4). Each of these will be introduced in a separate section below. In addition, we also include a section on methods using *distributional similarity* as a proxy for this task (Section 4.5), and *hybrid methods* (Section 4.6) that combine multiple purebred measures in certain ways. At the end of each section we present a graph illustration to summarize the methods and their connections (i.e., Figures 2–4, 8, 10, 11). Section 4.7 briefly summarizes this section and highlights the trend in recent research.

4.1 Path-based methods

The fundamental rationale behind path-based methods is that the relatedness between concepts can be determined based on their distance, or the length of the paths connecting them in a semantic graph following a given type of edges, or relation. The length of a path is typically calculated by counting the number of edges or nodes along the path. For this reason, they are also referred to as *edge-based methods* (Pesquita *et al.* 2009). The majority of these methods exploit taxonomic links in a semantic network; therefore, they measure similarity rather than relatedness.

The earliest work of this type is Rada *et al.* (1989), which measures semantic relatedness between two concepts using the shortest path length in a semantic graph. This method may be used to measure semantic similarity if edges of the path correspond to classical, taxonomic relations; or measure relatedness if nonclassical relations are accounted. When applied to taxonomic structures, the shortest path is effectively the one that connects the concepts by their *lcs*.

Despite the simplicity of this method, it has proved successful in their initial application to the MeSH concept hierarchy and in a number of follow-up studies using other knowledge bases such as Jarmasz and Szpakowicz (2003), Gentleman (2005), and Bhattacharya, Bhowmick and Singh (2010). However, Lee, Kim and Lee (1993) stated that the method simply based on the shortest path length can be unreliable, particularly when the paths are not restricted to the hypernymy relation.

Hirst and St-Onge (1998) suggested limiting the length of a valid path and discriminating the change of directions (*upward* such as hypernymy, *downward* such as hyponymy, and *horizontal* such as antonymy) along the path. The hypothesis is that the strength of relatedness correlates negatively with path length and frequency of changes of direction along the path. Yang and Powers (2005) proposed to calculate similarity as the product of edge weights rather than their sum. They also suggested that a valid path can contain different types (taxonomic relations) of edges, where most other path-based methods define a path as a consecutive set of edges of a particular type. The rationale is that different types of edges contribute differently to semantic similarity, where the weights are defined arbitrarily in their study.

Notions of specificity, depth, and density

A widely recognized limitation of these methods is that they do not account for the specificity of nodes in a taxonomy. Typically, edges at all levels in the taxonomy

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

are assumed to represent uniform length, and that nodes are distributed uniformly across the hierarchy. These assumptions are rarely true in practice, particularly for biomedical ontologies (Pesquita *et al.* 2009). This can be illustrated using Figure 1, where concepts at higher levels of a hierarchy tend to represent more general and abstract meanings whereas those lower down tend to represent more specific meanings. Thus, the same path length at different levels (e.g., the shortest path between *seafood* and *fruit*, and that between *spinach* and *chard*) denote different distances.

To overcome these issues, many have incorporated the notion of *depth* or *density*, or both to account for specificity. The motivation is that two nodes are semantically closer if they reside deeper in the hierarchy or are more densely connected locally. Thus, the measure of specificity should reflect the monotonic nature. Given a node c in a taxonomy T, the depth of c denoted by $depth(c)$ is usually the number of nodes along the longest path between c and *root*, and the depth of the taxonomy is the depth of the deepest node, or in mathematical terms, $depth(T) = max_{c \in T} depth(c)$. In Figure 1, $depth(spinach) = 4$, $depth(T) = 5$. The density of a node, denoted by $den(c)$, is usually defined as the number of its child nodes, or the number of its sibling nodes. In Figure 1, the density of *root vegetable* is 1 following the first definition, or 2 following the second definition.

These classic definitions of depth and density have been widely adopted. A recent study by Wang and Hirst (2011) has shown their limitations and proposed new measures for both. Using the WordNet ontology and several gold standard datasets, they carried out experiments to uncover the correlation between the depth or density of *lcs* of two words and their semantic similarity. In both cases, the correlation was found to be non-definitive. In terms of depth, the classic definition assumes a linear function that returns the depth of a node as an ordinal integer. However, experiments have shown that the 'notion of depth is relative to the distribution of number of nodes over depth value', and the distribution of nodes over depth conforms to a normal distribution. Also, there is 'no definitive, sufficient, and necessary relation between depth and similarity'. In particular, examples are given to show that semantically similar words (and their underlying concepts) are not necessarily deeper in the hierarchy. More experiments were carried out to show that the correlation between density (as by number of sibling nodes) and similarity is found to be even weaker. The distribution of density values is found to generally follow the Zipf's law, with more than 90% of nodes in WordNet having density values not greater than 3. This means that for the majority of concepts in WordNet, there are "only three integer values to distinguish the varying degrees of similarity."

Wang and Hirst (2011) therefore proposed new measures of depth and density that reflect their true distributional nature. These are denoted as $depth_{WH}$ and den_{WH}, respectively, and are calculated as below:

$$depth_{WH}(C) = \frac{\sum_{c' \in T} |\{c' : depth(c') \leq depth(c)\}|}{|T|} \quad (1)$$

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

$$den_{WH}(c) = \begin{cases} 0 & c = root \\ \dfrac{\sum_{c' \in ancestor(c)} den_{WH}(c')}{|ancestor(c)|} + den(c) & otherwise \end{cases} \quad (2)$$

In formula (1), $depth(\cdot)$ returns the depth of a concept based on the classic definition, and $|T|$ is the size of the taxonomy as the total number of nodes. Similarly in formula (2), $den(\cdot)$ returns the density of a concept following the definition based on sibling nodes. The authors then replaced the measures of depth and density in some previous methods with the new measures and obtained significant improvement in accuracy.

Due to the recency of this work, existing path-based methods still exploit the classic definition of depth and density. In the following, when we describe the depth or density of a concept, we always refer to their classic definitions, unless otherwise stated.

Addressing specificity in pathbased methods

Several studies (Ye *et al.* 2005; Yu *et al.* 2005; Lei and Dai 2006) have proposed methods that only exploit *depth* of concepts or their *lcs*. For example, Yu *et al.* (2005) defined similarity between two concepts as the ratio between their depths. Lei and Dai (2006) simply determined the similarity between two concepts as the depth of their *lcs*. Similar to purely path-based methods, these may also lead to spurious predictions, since they would tend to suggest concepts at the same level of a taxonomy (i.e., same depth) are similar, which is not necessarily true (e.g., following Lei and Dai (2006), pair-wise similarity between *spinach*, *chard*, and *cooking apple* tends to be similar using the taxonomy in Figure 1).

The majority of path-based methods combine *path length* with *depth* or *density*, or both in certain ways. Four widely cited methods in this direction are Sussna (1993), Wu and Palmer (1994), Jiang and Conrath (1997), and Leacock and Chodorow (1998), which have been thoroughly compared and discussed previously in Budanitsky and Hirst (2006). Leacock and Chodorow (1998) normalized the shortest path length between two nodes by the depth of the taxonomy; Wu and Palmer (1994) defined similarity between two nodes based on the depth of their *lcs* normalized with respect to the shortest path connecting them; Sussna (1993) and Jiang and Conrath (1997) combined both factors. In Sussna (1993), each edge connecting two nodes is assigned a weight, the calculation of which treats the edge as a combination of two unidirectional links, one leaving from a node to another and second as the inverse of this. The weight of each directed link is dependent on two factors: (1) the arbitrary weight for the relation it represents, and (2) the number of links of the same type leaving the same node (i.e., a measure of density). Then the path connecting two nodes is weighted following this method, and normalized by the depth of the two nodes. The method by Jiang and Conrath (1997) is usually classified as IC-based although the rationale is partly related to the path between concepts; this will be further discussed in Section 4.2.

In addition, Pekar and Staab (2002) and Liu, Zhou and Zheng (2007) also proposed methods that are highly similar to that of Wu and Palmer (1994), based

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

on similar rationales. According to Liu *et al.* (2007), the principle behind Wu and Palmer's (1994) method can be considered as computing semantic similarity based on the ratio of two concepts' common and different features, which can be quantified by the depth of their *lcs* and the shortest path length, respectively. Another study by Li *et al.* (2003) concluded with a measure that linearly combines path length with the depth of *lcs*, each assigned with a scaling factor to control the contribution of each.

Al-Mubaid and Nguyen (2006) introduced the notion of *common specificity*, which normalizes path length with respect to the depth of *lcs*. The common specificity of two concepts is determined by the depth of their *lcs*, and that of the cluster – essentially the branch, a sub-hierarchy – in the taxonomy that contains both concepts:

$$CSpec(c_1, c_2) = D - depth(lcs(c_1, c_2)) \tag{3}$$

where D is the depth of the cluster containing both concepts in T. The intuition is that 'the smaller the common specificity score, the more they share information, and thus the more they are similar'. Then let $\delta(c_1, c_2)$ denote the shortest path length between c_1 and c_2, their similarity is defined as a nonlinear combination of the shortest path length and common specificity:

$$SemSim(c_1, c_2) = log((\delta(c_1, c_2) - 1)^{\alpha} CSpec(c_1, c_2)^{\beta} + k) \tag{4}$$

where $\alpha > 0$ and $\beta > 0$ are contribution factors and $k \geq 1$ is a constant. Nonlinear functions are commonly used in semantic relatedness measures. Experiments have shown that nonlinear functions often give better prediction than linear functions (Li *et al.* 2003; Al-Mubaid and Nguyen 2006).

Wu *et al.* (2006) proposed a measure called *relative specificity similarity* that takes into account three different factors: the specificity of the *lcs* of two concepts in the taxonomy (α), the generality of each concept (β), and the local distance between the two concepts relative to their lcs (γ). The specificity of the *lcs* of two concepts is initially introduced in Wu *et al.* (2005) and is calculated as follows:

$$\alpha = max_{\substack{p(c_1,root)\in P(c_1,root) \\ p(c_2,root)\in P(c_2,root)}} \left\{ |concept(p(c_1, root)) \bigcap concept(p(c_2, root))| \right\} - 1 \tag{5}$$

where $p(c_1, root)$ denotes a single path from concept c_1 to *root*, $P(c_1, root)$ denotes the sets of available paths from c_1 to *root*, and $concept(p(c_1, root))$ is a function that returns the set of concepts or nodes along a path. For example, in Figure 1, $P(kelp, food)$ would return two paths: *kelp–sea vegetable–vegetable–food* and *kelp–sea vegetable–seafood–food*, where each path contains four concepts. Likewise there are two paths from *sea lettuce* to *food*, which together makes four pairs of paths for comparison. It is then straightforward to work out that the maximum intersection of these paired paths is three. This formula is specially tailored for taxonomies that allow multiple parents of a concept, which is typical in the biomedical domain. Although it is not explicitly defined with respect to *lcs*, effectively it counts the number of nodes along the path from $lcs(c_1, c_2)$ to *root* and thus has the equivalent effect of $depth(lcs)$. The generality factor β assumes that the similarity between two concepts is subject to the more general concept, which is determined based on their

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

position in the taxonomy. It is calculated in terms of path lengths, as

$$\beta = max\{min_{u \in leaf(c_1)}\{\delta(c_1, u)\}, min_{v \in leaf(c_2)}\{\delta(c_2, v)\}\} \tag{6}$$

where $leaf(c_1)$ returns the concepts that descend from c_1 (i.e., $descendant(c_1)$) and have no descendants of themselves (i.e., $descendant(u) = \oslash$, e.g., $leaf(fruit) = \{pear, bramley, jazz\}$). Then the local distance factor γ evaluates the distance between each concept to their shared *lcs*, and promotes concepts that are closer to the *lcs*:

$$\gamma = \delta(lcs(c_1), c_1) + \delta(lcs(c_2), c_2) \tag{7}$$

The final semantic similarity balances all three factors and the depth of the taxonomy, and is normalized into the scale between 0 and 1 as

$$SemSim(c_1, c_2) = \frac{depth(T)}{depth(T) + \gamma} \cdot \frac{\alpha}{\alpha + \beta} \tag{8}$$

The formula returns 0 when $\alpha = 0$, indicating that two concepts do not share any ancestors; and returns 1 when $\beta = 0$ and $\gamma = 0$, indicating that two concepts are the same leaf node.

Tsatsaronis, Varlamis and Vazirgiannis (2010) proposed a method that takes into account the weighted path length as well as the depth of all nodes along the path, which they call *path depth*. A major difference from other path-based methods is that a path can be a combination of different types of edges, including both taxonomic and non-taxonomic relations, while different types of relations are given different weights. The path length, which is captured under the notion of *compactness*, is computed as the product of weighted edges connecting the nodes along the path:

$$compactness(p(c_1, c_2)) = \prod_i^l w(e_i) \tag{9}$$

where $e_1, e_2 \ldots e_l$ are the edges along the path connecting c_1 and c_2, and $w(e_i)$ is a weighting function that assigns a real valued weight to an edge based on the type of relation it represents. The weights associated with each relation are designed to promote those that denote stronger semantic connections. The path depth, which is captured under the notion of *semantic path elaboration*, is calculated as the product of weighted depth of the nodes along the path:

$$spe(p(c_1, c_2)) = \prod_i^l \frac{2 \cdot depth(c_i') \cdot depth(c_{i+1}')}{depth(c_i') \cdot depth(c_{i+1}')} \cdot \frac{1}{depth(T)} \tag{10}$$

where $c_1', c_2' \ldots c_l'$ are the concept nodes along the path from c_1 to c_2. The intuition is to promote paths with deeper nodes, since paths with shallower nodes are more general. The final semantic relatedness between two concepts is the maximum product of *compactness* and *spe* given by any path between them.

While these path-based methods assume symmetric semantic relatedness, Schickel-Zuber and Faltings (2007) introduced a metric that allows measuring asymmetric similarity between concepts. The approach, called the *ontology structure-based similarity (OSS)*, firstly computes an *a-priori score* (APS) of every concept in a taxonomy to reflect its topological property; then views the similarity between two concepts as an effect of transferring the score from one concept to another via a directed path connecting the two concepts by their *lcs*. The APS of a concept is

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

calculated using formula (11), which has an inverse relationship with the number of descendants of a concept and therefore can be considered as a measure of density:

$$APS(c) = \frac{1}{descendant(c) + 2} \tag{11}$$

As mentioned before, the classic definition of *lcs* may return multiple concepts. The authors introduced a tie-breaking method to handle such situation, which balances the depth of the node with the number of different paths leading to the node from the two concept nodes in question:

$$lcs_{SF07}(c_1, c_2) = max_{c \in lcs(c_1,c_2)}\{(|P(c_1, c)| + |P(c_2, c)|) \cdot 2^{depth(c)}\} \tag{12}$$

The hypothesis is that 'a concept found higher in the ontology can still be more useful ... if it has many paths leading to it'. Following this definition, the single *lcs* connects the two concepts with a directed path, along which the score of the starting concept is transferred to the target concept. The amount transferred depends on an upward (α) transfer from c_1 to the *lcs*, and a downward (β) transfer from the *lcs* to c_2:

$$\alpha(c_1, lcs_{SF07}(c_1, c_2)) = APS(lcs_{SF07}(c_1, c_2))/APS(c_1) \tag{13}$$

$$\beta(c_2, lcs_{SF07}(c_1, c_2)) = APS(c_2) - APS(lcs_{SF07}(c_1, c_2)) \tag{14}$$

Intuitively, the upward transfer quantifies the amount of information of c_1 that can be generalized by the *lcs* as a ratio of their APS; whereas the downward transfer gives a sense of how much information becomes specialized by c_2. Next, based on the hypothesis that the distance between two concepts is correlated to the amount of transferred score, they introduced the following function to transform the transferred score into a normalized distance measure:

$$SemDist(c_1, c_2) = \frac{log(1 + 2\beta(c_2, lcs_{SF07}(c_1, c_2))) - log(\alpha(c_1, lcs_{SF07}(c_1, c_2)))}{maxD} \tag{15}$$

where *maxD* is the longest distance between any two concepts in the taxonomy.

Figure 2 summarizes path-based methods and the background information resources used in the original studies.

4.2 Information content-based methods

Information content-based methods hypothesize that the relatedness between concepts can be measured by the amount of information they share, which, if in a taxonomy, is often determined with respect to their *lcs*. These methods usually combine knowledge of a concept's hierarchical structure with statistics of its actual usage in text usually derived from a large corpus. The first IC-based method is introduced by Resnik (1995), which measures the IC of a concept as follows:

$$IC(c) = -log\, p(c) \tag{16}$$

where $p(c)$ is the probability of encountering an instance of a concept c, estimated from noun (which corresponds to the concept) frequencies observed in a large corpus. The counting ensures that the frequency of a noun is added to all of its

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

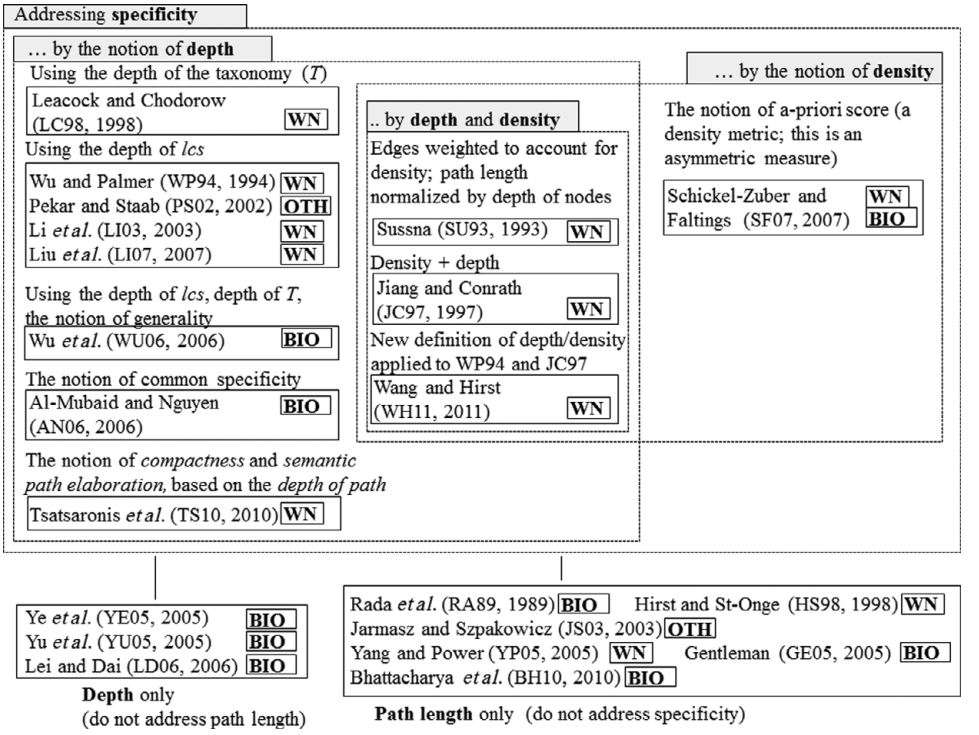


Fig. 2. A summary of path-based methods and the initial background information resources used. WN: WordNet, BIO: a biomedical knowledge base, OTH: other structured knowledge bases.

subsumers in the taxonomy, which guarantees p and therefore the IC of a concept, to be monotonic (i.e., $p(c) \leq p(c')$ and $IC(c) \geq IC(c')$ if c *IS-A* c'). The semantic similarity between concepts is assigned to be the maximum IC of their *cs*, which is effectively the *lcs*. Due to the monotonic nature of the IC measure, it can be considered as a measure of specificity of concepts (Li *et al.* 2003).

Resnik's (1995) definition of IC is widely adopted by later methods. Lord *et al.* (2003a) applied Resnik's method to the biomedical domain. Most of the later IC-based methods improved Resnik's in different ways to overcome a number of limitations. *The first limitation* is that any two pairs of concepts having the same *lcs* will receive the same similarity, which is not necessarily appropriate. Budanitsky and Hirst (2006) refers to this issue as 'indistinguishability'. Using the taxonomy in Figure 1, Resnik's method will output the same similarity for the pair *seafood–fruit* and *bramley–kelp*, which have the same *lcs, food*.

To address this problem, many methods incorporate the IC of each concept in question as a way to balance that of their *lcs*. Two widely cited studies in this direction are Jiang and Conrath (1997) and Lin (1998b). They have been summarized and compared in Budanitsky and Hirst (2006) and Zesch and Gurevych (2010a). Briefly, Jiang and Conrath (1997) proposed a measure based on the rationale of the shortest path but formulates the distance as a function of the IC measure. Lin's (1998b) semantic similarity measure aims to address the commonality of two concepts as

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

well as their difference, both measured in terms of IC. In its initial form the measure applies to taxonomic structures only. Maguitman *et al.* (2005) generalized Lin's measure such that it is applicable to both hierarchical and non-hierarchical links. A number of studies have adapted these methods to the biomedical domain with small modifications. For example, Speer, Spieth and Zell (2004) converted Lin's semantic similarity measure into a distance metric whereas Schlicker *et al.* (2006) combined Resnik's and Lin's methods to calculate similarity.

The second limitation with Resnik's definition of IC is related to the calculation of $p(c)$, which is dependent on the choice of a corpus. As a result of this, given different corpora, it is possible to obtain a different IC of a concept. Seco, Veale and Hayes (2004) argued that the IC of a concept should be related to its hierarchical structure, and introduced an alternative IC measure called 'intrinsic IC', calculated solely based on a taxonomy:

$$IC_{int}(c) = 1 - \frac{log(|descendent(c)| + 1)}{log(|T|)} \tag{17}$$

The rationale behind this method is that taxonomic structures usually organize concepts in such a way that the information expressed by a concept is inversely proportional to its descendants. Therefore, the IC value of a concept can be assessed using a function of the descendants it has. The purpose of the denominator is to normalize the IC value to the scale of 0 to 1. This notion of IC is adopted by Pirró (2009), who introduced a method based on the same hypothesis as that by Lin (1998b) to quantify semantic similarity between two concepts based on their commonality and difference:

$$\begin{aligned} SemSim(c_1, c_2) &= IC_{int}(lsc(c_1, c_2)) - (IC_{int}(c_1) - IC_{int}(lsc(c_1, c_2))) \\ &\quad - (IC_{int}(c_2) - IC_{int}(lsc(c_1, c_2))) \\ &= 3 \cdot IC_{int}(lsc(c_1, c_2)) - IC_{int}(c_1) - IC_{int}(c_2) \end{aligned} \tag{18}$$

where $IC_{inst}(lsc(c_1, c_2))$ represents the commonality of two concepts and $IC(c) - IC(lsc)$ represents the features that are unique to the concept.

While these methods generally define semantic similarity of two concepts with respect to their *lcs*, which is usually a single concept, some methods have proposed to consider *multiple common subsumers* of the concepts. Couto, Silva and Coutinho (2005) proposed the GraSM measure, which takes the average of IC of all *common disjunctive subsumer* ($cs_{disjunct}$) of two concepts:

$$SemSim(c_1, c_2) = \frac{\sum_{c \in cs_{disjunct}(c_1, c_2)} IC(c)}{|\{c | c \in cs_{disjunct(c_1, c_2)}\}|} \tag{19}$$

Two common subsumers are said to be *disjunctive* if there are 'independent paths from both ancestors to the concept', where an independent path is one that contains at least one concept unused by other paths. The motivation is that each *common disjunctive subsumer* provides a different interpretation of concepts that can be equally important. For example, with an *lcs*-based method, the similarity between *leaf vegetable* and *sea vegetable* in the taxonomy of Figure 1 depends on their *lcs, vegetable*. GraSM also takes into account the following independent paths to

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

food: *leaf vegetable–vegetable–food* and *sea vegetable–seafood–food*, and thus also considers the IC of *food*. On the other hand, there are no independent paths from *root vegetable* and *leaf vegetable* to *food*, and therefore their similarity is only dependent on their *lcs, vegetable*. Thus, following the formula, GraSM will return a higher similarity for *root vegetable* and *leaf vegetable* than *leaf vegetable* and *sea vegetable*. The intuition is that *sea vegetable* has another interpretation (*seafood*).

Wang *et al.* (2007) also argued that firstly all ancestors of a concept should contribute to the semantics of the concept, and therefore should be accounted for when measuring semantic similarities between concepts; secondly, the significance of the contribution should follow an inverse relationship with their distance to the concept. Given a hierarchical structure containing a concept c, their method firstly extracts a sub-hierarchy to represent the semantics of c. Formally, $T_c(c, C_c, E_c)$ is the extracted sub-hierarchy for the concept c, where C_c denotes all concepts in the sub-hierarchy, including *ancestor*(c) and c itself, and E_c denotes the set of edges that connect any two concepts in the sub-hierarchy. Thus, each member concept in C_c contributes to the semantics of c, via a path that connects the two concepts by edges in E_c. This contribution is quantified by

$$semantics_c(c') = \begin{cases} 1 & c = c' \\ max_{ch \in children(c'), e \in E_c} \{weight(e) \cdot semantics_c(ch)\} & c \neq c' \end{cases} \quad (20)$$

where $semantics_c(c')$ is interpreted as the contribution of semantics of c' to c. The formula states that the contribution of semantics by the concept itself ($c = c'$) is 1, since intuitively they are identical; while the contribution by any ancestor of c ($c \neq c'$) is calculated based on the contribution of its child concepts, along weighted edges that connect itself with its children. The weight allows discriminating different semantic relations (e.g., *IS-A, PART-OF*) of edges. Next, the contribution by all members in C_c is aggregated to derive a measure *semantic value* of c:

$$sv(c) = \sum_{c' \in C_c} semantics_c(c') \quad (21)$$

The final semantic similarity between two concepts is calculated by their 'semantics' in common, which follows similar rationale with other IC-based methods:

$$SemSim(c_1, c_2) = \frac{\sum_{c \in C_{c_1} \cap C_{c_2}} (semantics_{c_1}(c) + semantics_{c_2}(c))}{sv(c_1) + sv(c_2)} \quad (22)$$

where $C_{c_1} \cap C_{c_2}$ is equivalent to $cs(c_1, c_2)$.

A summary of IC-based methods is shown in Figure 3.

4.3 Gloss-based methods

As introduced previously in Section 3.1, concepts represented in knowledge bases are often provided with definitions and examples. Gloss-based methods generally refer to these as **glosses**, and hypothesize that the relationship between concepts is often implied by the shared words in their glosses. Therefore, gloss-based methods

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

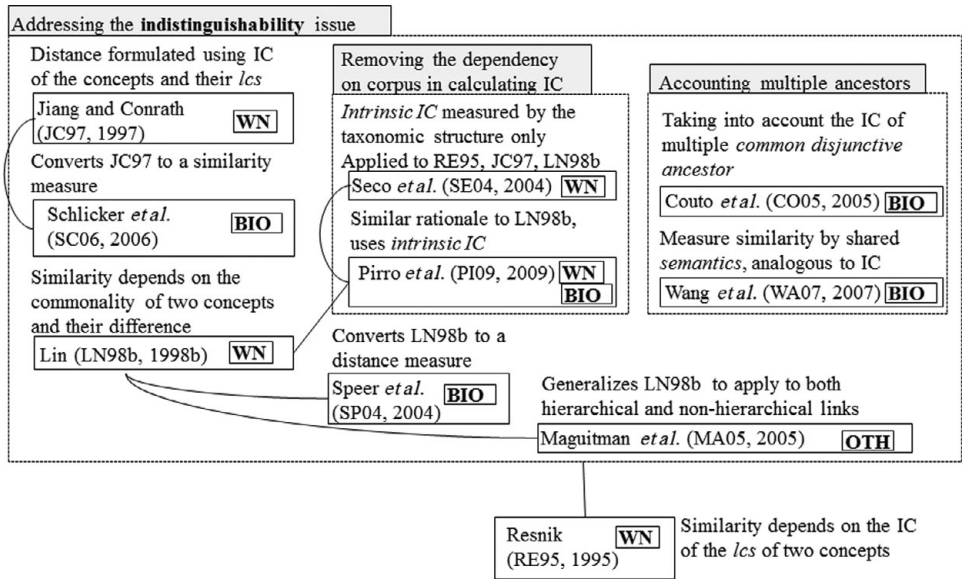


Fig. 3. A summary of IC-based methods and the initial background information resources used. WN: WordNet, BIO: a biomedical knowledge base, OTH: other structured knowledge bases.

propose to measure semantic relatedness with respect to the *overlap of words* in two concepts' glosses. Following this definition, gloss-based methods assess relatedness rather than similarity, since they do not make use of taxonomic links. However, there are also exceptions.

Lesk (1986) introduced the first gloss-based measure, which simply evaluates the relatedness between concepts by the number of overlapping words in their glosses, namely,

$$SemRel(c_1, c_2) = \left| gloss(c_1) \bigcap gloss(c_2) \right| \tag{23}$$

where $gloss(c)$ returns the set of words extracted from the gloss of c. This simple measure has been adopted by several later studies, including Mihalcea and Moldovan (1999), Banerjee and Pedersen (2003), and Gurevych (2005), which are distinguished by the construction of $gloss(c)$. For example, Banerjee and Pedersen (2003) argued that the gloss of a concept should be extended by including the glosses of others that are related to the concept; while Gurevych (2005) adapted the method to knowledge bases that do not provide a gloss of a concept by building a pseudo-gloss, which concatenates concepts in close relation (e.g., hypernym, synonym) to the concept. A detailed survey of these methods can be found in Zesch and Gurevych (2010a).

Further to the motivation behind Gurevych (2005), if we lessen the definition of gloss to allow different methods of gloss construction, a few other methods can also be classified as gloss-based. Gentleman (2005) proposed to measure the overlap of ancestors of two concepts extracted from a taxonomic structure, where *ancestor*(c)

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

can be considered as a gloss representation of c,

$$SemSim(c_1, c_2) = \frac{|ancestor(c_1) \bigcap ancestor(c_2)|}{|ancestor(c_1) \bigcup ancestor(c_2)|} \tag{24}$$

This method can be considered to measure similarity because it uses ancestors of concepts only. The general hypothesis is that each ancestor of a concept provides an interpretation, and the numerator is the amount of interpretations common to both concepts, while the denominator is the amount required to fully interpret both concepts. The similarity can be evaluated as the ratio of two values. Turdakov and Velikhov (2008) adopted the Dice measure, which is in a very similar form. The pseudo-gloss are constructed using Wikipedia article links. Briefly, a pseudo-gloss for each concept is constructed by concatenating the outgoing and incoming links of the article page describing that concept, and placed into the same formula.

While the method by Gentleman (2005) can be considered as quantifying similarity with respect to the shared knowledge of two concepts, Batet, Sánchez and Valls (2011) proposed a method that quantifies similarity with respect to 'non-shared knowledge'. The method represents each concept as a gloss of their ancestors in the same way, and takes into account the 'non-shared' and 'shared knowledge' between two concepts, which is calculated based on the overlap of their gloss. Let $G(c)$ be the set containing $ancestor(c)$ and c, then the similarity is defined as

$$SemSim(c_1, c_2) = -\log_2 \frac{|G(c_1) \bigcup G(c_2)| - |G(c_1) \bigcap G(c_2)|}{|G(c_1) \bigcup G(c_2)|} \tag{25}$$

To some extent, the rationale behind these methods (Gentleman 2005; Turdakov and Velikhov 2008; Batet *et al.* 2011) is related to that of IC-based methods, in the sense that relatedness is quantified based on the information that two concepts share in common. However, they are classified as gloss-based methods rather than IC-based for two reasons: (1) they do not quantify the information content of a concept formally; and (2) the calculation is generally based on certain forms of vocabulary overlap, which is the key feature of gloss-based methods.

A summary of gloss-based methods is shown in Figure 4.

4.4 Vector-based methods

This section discusses vector-based methods, which according to a narrow definition by Zesch and Guryvech (2010a), refer to methods that represent a term or concept using a feature vector derived from a structured knowledge base, rather than using co-occurrence counts or contexts. Feature vectors are widely used for data representation. For vector-based methods, concepts and their lexicalized forms can be represented based on features encoded in knowledge bases. For example, in the taxonomy shown in Figure 1, concepts can be described by the features *parent concepts* and *child concepts*. Figure 5 illustrates the vector representations of *vegetable* and *seafood* based on these two types of features.

With feature vector representations, assessing relatedness between concepts can be achieved by comparing the similarity between their feature vectors. The most well-known and widely used measure for this purpose is the cosine similarity function,

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

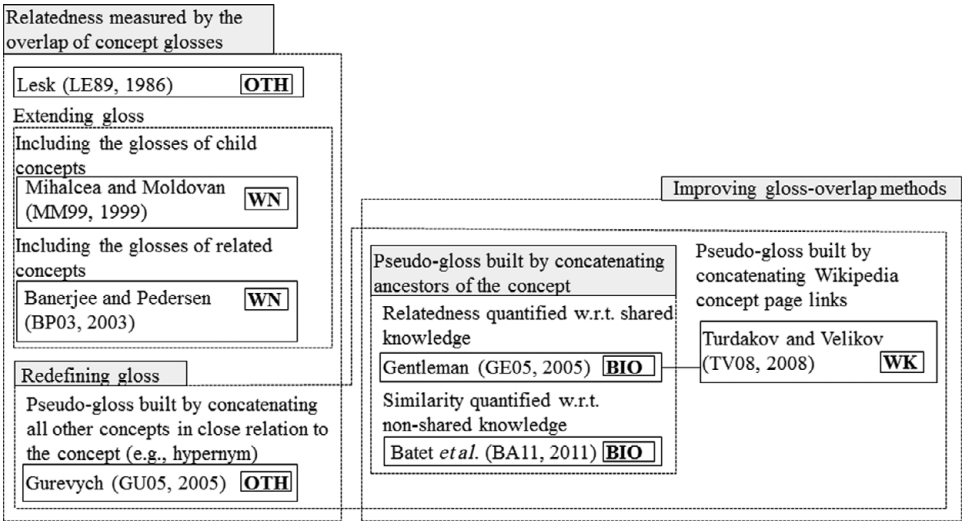


Fig. 4. A summary of gloss-based methods and the initial background information resources used. WN: WordNet, WK: Wikipedia, BIO: a biomedical knowledge base, OTH: other knowledge bases.

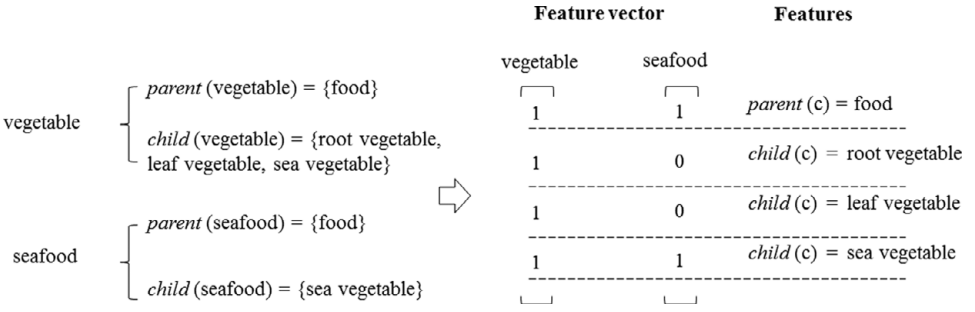


Fig. 5. Feature representations of concepts based on the taxonomy in Figure 1. 0 – a feature is inactive for the concept; 1 – a feature is active for the concept.

which measures the similarity between two vectors by the cosine of the angle between them:

$$cosine(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1, \cdot \vec{v}_2}{||\vec{v}_1|| \cdot ||\vec{v}_2||} \tag{26}$$

The cosine similarity measure is used by the majority of vector-based methods. However, they differ significantly in terms of how the feature vectors are constructed. Because such methods do not exploit the hierarchical nature of a knowledge base, they generally quantify relatedness rather than similarity.

Generally, we classify vector-based methods into two types. The first type of methods build a concept vector using the lexical semantic content (e.g., a concept's gloss and synonyms in WordNet) directly defined for that concept in a knowledge base. Intuitively the lexical semantic content can be considered as features relevant to a concept. The second type of methods builds a concept vector using other

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

related concepts in the knowledge base. The intuition is that semantically related concepts are related to a similar set of concepts. In these methods, features for each relevant concept are exploited indirectly and usually contribute to the weight of the relevant concept in the corresponding vector of the target concept. The classification is illustrated in Figure 8 at the end of this section.

Patwardhan and Pedersen (2006) represented a concept by a second-order gloss vector using WordNet glosses. Firstly, WordNet is turned into a corpus made of the set of glosses in WordNet. Next, a first-order context vector based on co-occurrence is created for every word in this corpus, where words are said to co-occur if they are found in the same gloss. Then for each concept, a second-order gloss vector is constructed by combining the first-order context vectors of words that appear in its gloss. For example, the second-order gloss vector for the concept *cat (mammal)* is built using the resultant first-order vectors of nonstop words (shown here in *italic*) in its WordNet gloss '*feline mammal* usually having *thick soft fur* and no *ability* to *roar*'. The intuition is that the orientation of the second-order gloss vector indicates the domains or topics it is associated with. The relatedness between two concepts is then measured as the cosine similarity between two vectors.

Zesch *et al.* (2008b) suggested that different kinds of lexical semantic content defined for a concept can be used as features to build concept vectors. They demonstrated this using three different knowledge bases with the generic cosine similarity function. Concept vectors are built using the first paragraph of Wikipedia, WordNet gloss, and Wiktionary gloss separately. Zhang *et al.* (2011) proposed to combine various kinds of lexical semantic content from different knowledge bases to create joint vector representations. The hypothesis is that (1) a concept may be covered in different knowledge bases, and (2) each knowledge base provides different kinds of lexical semantic content (e.g., the glosses, synonyms, and hypernyms in WordNet; the article content, redirects, and categories in Wikipedia) that present complementary perspectives to each other. To exploit this, the authors proposed to build a combined vector representation of concepts in three steps. Firstly, given a polysemous term and its corresponding concept articles found in Wikipedia and all its synsets in WordNet, a Wikipedia concept article is mapped with a WordNet synset that is likely to refer to the same meaning. This is done by using a simple gloss overlap-based method. Secondly, different kinds of lexical semantic content in each knowledge base are cross-mapped following the sense of similar *semantics*. For example, frequent words used in a Wikipedia article are joined with WordNet gloss; Wikipedia redirects are joined with WordNet synonyms; and the categories assigned to the article are joined with WordNet hypernyms. Finally, a joint feature vector representation is created based on the cross-mapped content from both knowledge bases. Szarvas, Zesch and Gurevych (2011) also proposed to harness diverse knowledge bases, including Wikipedia, WordNet, and Wiktionary. The method firstly represents a concept using features extracted from each knowledge base separately and then computes three different scores. The scores are then used to train a model in a supervised setting.

One of the methods that represent a word or concept as a vector of related concepts is introduced in Ziegler, Simon and Lausen (2006). They firstly queried

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

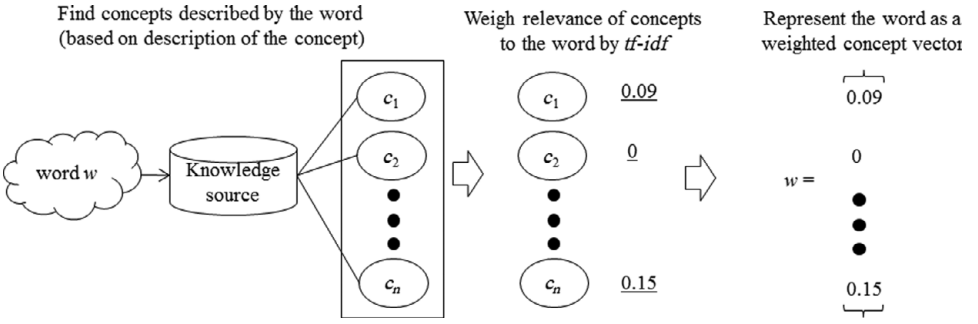


Fig. 6. Illustration of the word representation process in ESA.

each word on Google Directory,[7] which returns a ranked list of websites each annotated with a category defined in the Open Directory Project (ODP) taxonomy. The ODP taxonomy organizes category concepts in a tree structure, and classifies websites using the categories. The word is then represented as a vector $\vec{v}$ of ODP categories, such that $|\vec{v}| = |T|$ the size of the ODP taxonomy. The initial weight of each category *cat* in the vector is dependent on the rank of the website that the corresponding category is associated with. Next, for each *cat*, its weight is propagated upwards to the ancestors of *cat* along the ODP taxonomic links, where the amount of propagation depends on the density of the parent node at each step. This process ensures that each ancestor of a *cat* receives a weight that is monotonically decreasing with respect to their depth. Then the final weight of *cat* is adjusted to combine the weights of all of its ancestors, resulting in the final weighted vector representation of the word. The intuition behind this is that a category could be counted as an instance of its ancestors, and therefore its weight should reflect the contribution from all its ancestors. The final semantic relatedness between two words is computed as the correlation between two vectors using the Pearson's correlation coefficient:

$$\begin{aligned} SemRel(w_1, w_2) &= SemRel(\vec{v}_{w_1}, \vec{v}_{w_2}) \\ &= \frac{\sum_{k=0}^{|T|}(v_{w_1,k} - \bar{v}_{w_1}) \cdot (v_{w_2,k} - \bar{v}_{w_2})}{\left(\sum_{k=0}^{|T|}(v_{w_1,k} - \bar{v}_{w_1})^2 \sum_{k=0}^{|T|}(v_{w_2,k} - \bar{v}_{w_2})^2\right)^{\frac{1}{2}}} \end{aligned} \tag{27}$$

where $\bar{v}_x$ is the mean value of $\vec{v}_x$, and T refers to the ODP taxonomy.

A similar approach is Explicit Semantic Analysis (ESA) introduced by Gabrilovich and Markovitch (2007). ESA computes word relatedness using vectors constructed using Wikipedia concepts. The method firstly builds an inverted index of words and Wikipedia articles, then represents each word as a high-dimensional vector of Wikipedia articles (Figure 6). Since each article usually focuses on certain topic, the vector can be viewed as a vector of concepts, where each element in the vector

[7] Originally http://www.google.com/dirhp, which has been closed

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

corresponds to a concept and the dimension is the number of Wikipedia articles. Each element in the vector is weighted by the *tf.idf* (Jones 1973) score of the word in the associated article. The intuition is that the weight denotes the degree of relevance between the word and the concept. Semantic relatedness is also calculated using the cosine function.

Explicit semantic analysis was later extended by Hassan and Mihalcea (2009) and Radinsky *et al.* (2011). Hassan and Mihalcea (2009) introduced two modifications to address corresponding issues in the original ESA method. Firstly, the weight of each element in the vector is normalized to account for the length of the associated concept article, as the original method may be biased toward articles with lengthy descriptions. Secondly, the weights for the vector elements corresponding to a concept found in the Wikipedia category tree are scaled according to their depth in the category tree. The intuition is to promote concepts that are lower down (and thus more specific) in the category tree. They also replaced the cosine similarity metric with an overlap-based metric similar to that by Lesk (1986) in order to place more emphasis on the overlap of vectors.

Radinsky *et al.* (2011) proposed to incorporate the 'temporal behavior' of words in computing their relatedness. The idea originates from the observation that semantically related words do not necessarily co-occur in the same articles; however, they are likely to be mentioned roughly around the same time. For example, by studying the distributional patterns of words over time using a collection of *New York Times* articles spanning over 130 years, they found that the words *war* and *peace* tend to correlate in their usage frequency over time. However, they might rarely be mentioned at the same time in the same articles. To exploit this feature they proposed to modify the ESA method by modelling the 'temporal dynamics' of each non-zero weighted concept in the vector, and scaling their temporal dynamics according to the concept's original weight in the vector. The temporal dynamics of a concept is modelled as a series of its usage frequency over the corpora:

$$dynamics(c) = \left\langle \frac{\{d \in D_1 | freq(c,d)|)\}}{|D_1|}, \ldots, \frac{\{d \in D_n | freq(c,d)|)\}}{|D_n|} \right\rangle \tag{28}$$

where $D_1 \ldots D_n$ can be viewed as a history represented by a collection of corpora ordered chronologically, D_i is a corpus at time point I, and $freq(c,d)$ returns the frequency of observing the lexicalized words of the concept in d. Details of the counting method can be found in Radinsky *et al.* (2011). Next, each concept in the ESA vector is represented by its temporal dynamics, the weight of which inherits that of the concept in the original ESA vector. The final relatedness between the two words can be computed as the sum of pairwise concept relatedness of non-zero weighted concepts in their vectors. And the pairwise concept relatedness is computed using their temporal dynamics by two different methods for measuring time series similarity.

Milne and Witten (2008) employed the hyperlinked Wikipedia article graph and proposed to represent a Wikipedia concept as a vector of concepts linking to it or a vector of concepts it links. The vector in the first case is made of incoming links in a concept article, while in the second case it is made of outgoing links. The

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

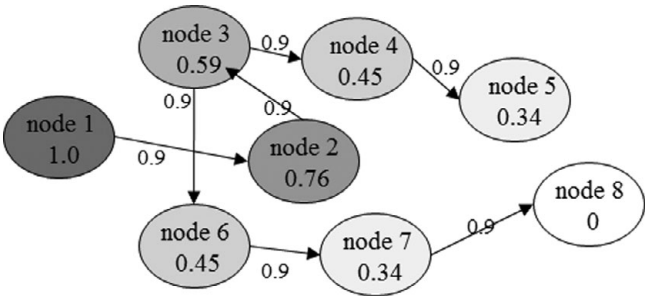


Fig. 7. A spreading activation network with a decay factor of 0.9 and firing threshold of 0.3. The initial activation value of 1.0 is propagated from node 1 through the edges.

cosine similarity function and the measure proposed by Cilibrasi and Vitanyi (2007) were applied to each vector representation, respectively, for evaluation. Cilibrasi and Vitanyi's (2007) method will be described in Section 4.5, since initially it was developed as a distributional similarity measure. Liu and Chen (2010) also represented Wikipedia concept as a vector of links, but defined the relatedness between two concept vectors as the sum of the pairwise relatedness between any two links across two vectors, which is then computed using either gloss overlap, or the Wu and Palmer's (1994) measure using the category graph.

The method by Milne and Witten (2008) only considers the adjacent concepts (e.g., neighbors) in a semantic graph. Another stream of work based on the link structure of a semantic network exploit both direct and indirect connections between concepts. Gouws, van Rooyen and Engelbrecht (2010) proposed to apply the spreading activation algorithm (Collins and Loftus 1975) to a semantic network of concepts as a method to derive concept vectors.

Briefly, given a semantic network of nodes connected by directed and weighted edges, such as shown in Figure 7, spreading activation allows propagating the *activation value* from a given node (or a set of nodes) to any other nodes in the network across the path that connects them. For each pair of adjacent nodes connected by a directed edge along the path, the propagation is controlled by two factors: a *firing threshold* that checks whether the activation value of the current node should be passed on – thus firing the propagation – to its adjacent nodes; a *decay factor* decides the amount of value passed on and accumulated if the propagation is fired. The intuition is that as the propagation travels a long distance, the value propagated drops. For the purpose of measuring semantic relatedness between two concepts c_1 and c_2, Gouws *et al.* (2010) built a semantic network of Wikipedia concepts connected by incoming links from the article page. They firstly set a non-zero initial activation value to the node of concept c_1, while all other nodes receive an initial value of 0. Spreading activation is started to propagate the value to c_2 through each node along the set of paths connecting them. After the activation terminates (determined by a number of methods), c_2 receives the final activation value, denoted by $act(c_2|c_1)$. Meanwhile, a concept vector $\vec{v}_1$ is created for c_1 by collecting the final activation values of all nodes in the network. These values can be considered as a measure of the relevance between each node and c_1. Next, the same

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

process is repeated from the opposite direction, by propagating an initial activation value from c_2 to c_1, to obtain $act(c_1|c_2)$ and $\vec{v}_2$. Finally, the authors computed the semantic relatedness between the two concepts in three different ways: (1) as a function of the sum of $act(c_2|c_1)$ and $act(c_1|c_2)$, in which each value is considered to be a distance from one to another; (2) as the cosine similarity of two vectors; and (3) using a distance-based formula introduced in Cilibrasi and Vitanyi (2007).

Harrington (2010) and Wojtinnek and Pulman (2011) applied the same technique but derived a semantic and syntactic structure from an unstructured corpus by applying NLP techniques such as Named Entity Recognition and syntactic parsing to a corpus to generate a syntactic and semantic network of entities and concepts. The authors argued that the advantage of using a corpus is that they may provide better coverage of domain-specific information than a general-purpose knowledge base.

Another popular stream of methods that exploit the link structure of a semantic network for constructing concept vectors is using Personalized PageRank (Haveliwala 2002), an extension of the well-known PageRank technique. PageRank was initially an algorithm for calculating the importance of webpages by exploiting the hyperlinks between them. It computes a stationary probability distribution that represents the chance that a person randomly clicking on links will arrive at any particular page given the entire collection, regardless of which webpage to start from. Haveliwala (2002) extended this technique to allow the ranking (or importance) of each page to be biased (or 'personalized') toward a particular query, such that webpages more relevant to the query receive higher importance and therefore higher ranks. Such ideas have been adapted to a semantic network of concepts to develop semantic relatedness methods (Hughes and Ramage 2007; Yeh *et al.* 2009; Yazdani and Popescu-Belis 2010). Generally, the PageRank algorithm is applied to the semantic network of concepts to obtain the stationary probability distribution to represent the likelihood of reaching any node in the network. The final probability distribution depends on the connectivity of the network and weights of edges, but is independent on the starting node. Next, this distribution is personalized against each question concept node following Haveliwala's method, which effectively places more weights onto nodes closer to the target concept in the graph. Intuitively, this can be viewed as re-ranking all nodes (similar to webpages) in the network (similar to the web) with respect to their connection (similar to relevance) with the target concept node (similar to a query). The two resultant distributions, biased toward c_1 and c_2 respectively can be represented as two vectors, the dimensionality of which is the number of nodes in the network. Then the relatedness between them can be computed with a measure of vector similarity.

A major difference between these methods is how the semantic network is constructed. Hughes and Ramage (2007) constructed a graph of WordNet synsets, tokens (e.g., polysemous words), and token-with-part-of-speech (tokenPOS, e.g., click#noun, click#verb). Edges are established for any relations between synsets defined in WordNet, for a synset node, and a tokenPOS node if the synset 'uses' the tokenPOS (e.g., synset 'click#noun#mouse click' uses 'click#noun'), and for a tokenPOS node with a token node following the same strategy. Yeh *et al.* (2009)

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

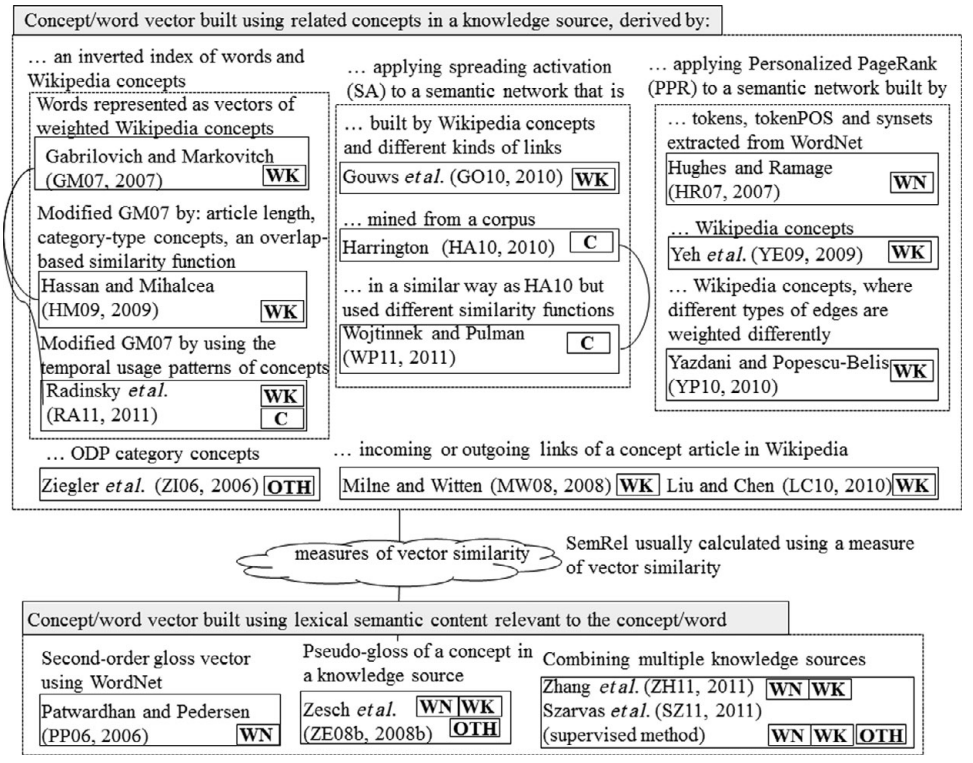


Fig. 8. A summary of vector-based methods and the initial background information resources used. WN: WordNet, WK: Wikipedia, OTH: other knowledge bases, C: a general-purpose corpus.

and Yazdani and Popescu-Belis (2010) constructed a graph of Wikipedia concepts. An edge is established between two concepts if they are connected by links in their article page by sharing a category, or sharing a link in their infoboxes. Yazdani and Popescu-Belis (2010) assigned higher weights to certain types of edges to gain better performance.

A summary of vector-based methods is shown in Figure 8.

4.5 Distributional similarity as a proxy for lexical semantic relatedness

A loose definition of distributional similarity based on the famous statement by Firth (1957), 'You shall know a word by the company it keeps', says that two words are similar if they appear in similar contexts. Weeds (2003) summarized the literature and defined that two words are distributionally similar if (1) they tend to occur in each other's context; or (2) the contexts each word tends to occur in are similar; or (3) if one word is substituted for another in a context, its 'plausibility' is unchanged. Different methods have adopted different definitions of contexts, but usually a context is the set of words collected from the window around the word, or an entire document, or syntactic relationship as introduced in Lin (1998a) and Weeds (2003). Following this definition, a multitude of methods are available for

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

measuring distributional similarity. These are not the focus of this survey; however, details can be found in Weeds (2003) and Turney and Pantel (2010).

There is a widely accepted hypothesis that distributional similarity can predict semantic similarity and thus relatedness in general. As a result, many studies have proposed to use distributional similarity methods as a proxy for lexical semantic relatedness. However, Budanitsky and Hirst (2006) argued that there are three essential differences between the two paradigms. Firstly, semantic relatedness is inherently a relation on concepts, while distributional similarity is a relation on words; secondly, semantic relatedness is typically symmetric whereas distributional similarity can be potentially asymmetric; finally, semantic relatedness depends on a structured lexicographic or knowledge bases, distributional similarity is relative to a corpus. Nevertheless, because of the popularity of their application to semantic relatedness, in the following we discuss several recent studies that apply distributional similarity methods to such tasks. Our criteria for selection is that the method is either evaluated using methods or datasets (see Section 5) usually used for evaluating lexical semantic relatedness, or that it is later ported to address relevant tasks in other studies. A popular research direction taken by these approaches is exploiting the Web as the background information corpus.

Matsuo *et al.* (2006) computed distributional similarity of two words using statistics collected from search engines. The method queries each word using a search engine to retrieve the number of pages containing each word denoted by $freq(w)$. Next, the number of pages containing both words $freq(w_1, w_2)$ is counted as the page counts for a query concatenating both words. Then using these figures, the distributional similarity can be calculated using the Point-wise Mutual Information (PMI) measure or the Chi-square (x^2) statistical test. PMI measures the strength of association between two words taking into account the probability of encountering each word in the corpus and the probability of encountering both together. While Chi-square also takes into account the probability of encountering one word but not another, as well as the probability of encountering none of the words, essentially, in Matsuo *et al.*'s (2006) method, the context is equivalent to a webpage.

Cilibrasi and Vitanyi (2007) introduced the Normalised Google Distance (NGD) measure, which builds on a data compression-related theory that is rather intricate. The resultant method is, however, very simple and also utilizes page counts:

$$SemDist(w_1, w_2) = \frac{max\{log(freq(w_1)), log(freq(w_2))\} - log(freq(w_1, w_2))}{logN - min\{log(freq(w_1)), log(freq(w_2))\}} \quad (29)$$

As mentioned before, this method is adopted by Milne and Witten (2008) for measuring semantic relatedness between concepts using Wikipedia. Briefly, they substituted $freq(w)$ as the number of incoming links leading to a concept article and $freq(w_1, w_2)$ as the number of incoming links leading to both concept articles.

While these methods based on page counts can be effective, a potential issue is that they do not consider the relative position of words or multiple occurrences of the word in a single page. Two words occurring in a page may not be related at all if they are distant. Therefore, methods based on page counts are prone to errors (Bollegala *et al.* 2007).

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Many methods cope with this by using the returned snippets for the query. Chen *et al.* (2006) proposed to count word co-occurrences in the top N snippets returned by a search engine. They hypothesize that two words w_1 and w_2 are associated if it is possible to find w_2 from w_1 (a forward process) and find w_1 from w_2 (a backward process) by web search. Therefore, their method queries each word in turn, and count the occurrences of the other in the top N snippets returned for the query (i.e., $freq(w_1|w_2)$ and $freq(w_2|w_1)$). This process gives two different co-occurrence figures for the two words, which they call a 'double checking' process. The word similarity is then computed using a Co-Occurrence Double Check (CODC) function as shown below:

$$DistSim(w_1, w_2) = \begin{cases} 0 & \text{if } freq(w_1|w_2) = 0 \\ & \text{or } freq(w_2|w_1) = 0 \\ e^{log(\frac{freq(w_2|w_1)}{freq(w_1)} \cdot \frac{freq(w_1|w_2)}{freq(w_2)})^{\alpha}} & \text{otherwise} \end{cases} \tag{30}$$

where α is a control parameter. In addition, a number of classic functions for assessing association strength such as Dice, Jaccard, and Cosine are also tested using the co-occurrence data. Sahami and Heilman (2006) queried each word to obtain a set of snippets. For each word, each of its corresponding snippets is represented as a weighted term vector, and the centroid of the set of vectors is computed. Similarity between the two words is defined as the inner product between the corresponding centroid vectors.

Although these methods can tackle the limitation of page counting-based methods to certain extent, one of their limitations, as suggested by Ruiz-Casado, Alfonseca and Castells (2005), is that they ignore word order and phrasal structures. The authors thus proposed a method that assesses the substitutability of two words as a measure of their similarity. It firstly collects a set of sentences (S_1, S_2) from the snippets returned for each word (w_1, w_2) as a query to a search engine. Next, it counts in how many of the sentences in S_1 it is possible to substitute w_1 with w_2. This is done by replacing w_1 with w_2 to create a new sentence, which is queried using the search engine for validation. The same process is repeated for S_2 and the final similarity score is derived based on the percentage of sentences that are substitutable.

Bollegala *et al.* (2007) proposed a method that combines both page counts and sentence-level contexts in snippets. Their method is based on supervised classification trained on examples, which is uncommon in semantic relatedness and distributional similarity methods. A Support Vector Machine (SVM)-based (Vapnik 1998) classifier is trained using a set of synonym pairs as positive examples and a set of non-synonym pairs as negative examples, both of which are randomly selected from WordNet. Part of the process is illustrated in Figure 9.

Firstly, a pattern induction process is applied to extract lexical patterns that are likely to indicate synonyms in a text. Using the positive and negative examples, each pair is queried to obtain a set of snippets containing both words. Then from each snippet, they extract n-gram lexical patterns from the context windows that include both words. However, many patterns may be found for both positive and negative

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

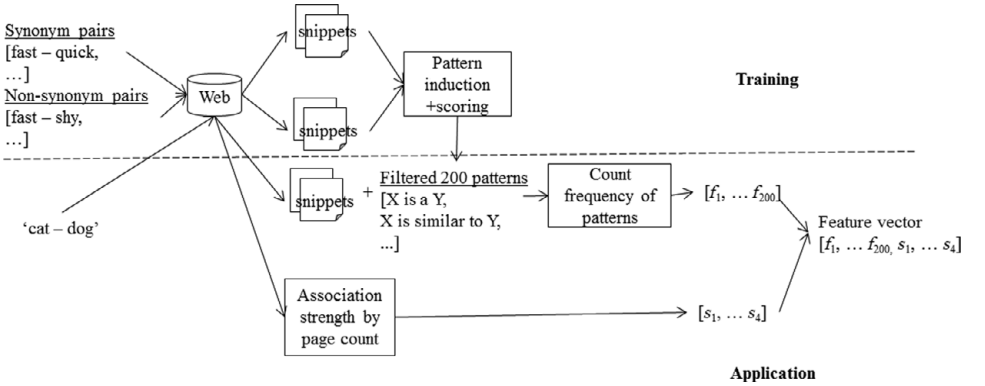


Fig. 9. Illustration of the method by Bollegala *et al.* (2007). Lexical patterns are extracted for synonym and non-synonym pairs (training data) based on search results from the Web. The patterns are filtered and used as features to represent the training data, using which a model is learnt. It is then used to classify new term pairs, based on the same feature representation model.

examples. Thus, for each pattern they count its frequencies of being found with positive and negative examples respectively (positive and negative frequencies), and compute a confidence score using the Chi-square function. The patterns are ranked by this score and only the top 200 patterns are selected. To train the classifier, they generate a feature vector for each of the positive and negative example. A pair of words is queried using a search engine, and the frequency of each of the 200 patterns is counted in the returned snippets. Next, four different functions of association strength are applied to compute four scores for the two words using page counts in a similar way to Matsuo *et al.* (2006). A feature vector is then created for the pair by concatenating the frequencies collected for 200 patterns and four scores. The SVM-based classifier is then trained using this training data to create a classification model, which is then able to predict similarity between any new pairs of words given a feature representation created in the same way.

A summary of distributional similarity methods discussed above is shown in Figure 10.

4.6 Hybrid methods

We refer to hybrid methods as those that combine multiple semantic relatedness methods or distributional similarity methods to arrive at a single measure of semantic relatedness.

Many of these methods firstly calculate semantic relatedness using an assembly of different purebred methods, and then derive an aggregated score as the average, sum or maximum of all scores. This kind of simple *combination* is intuitive – if each distinctive method provides a different perspective of semantic relatedness, it is natural to combine them to create a full picture. For example, Alvarez and Lim (2007) calculated semantic relatedness as the maximum score given by either a method that inverts the shortest path length (weighted by depth of concepts

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

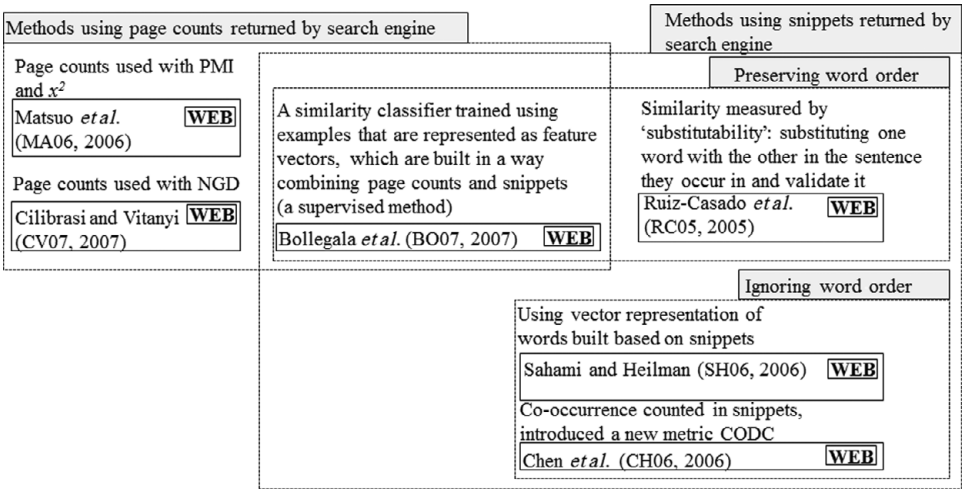


Fig. 10. A summary of distributional similarity methods. All methods used the web (WEB) as the background information resource.

along the path), or a method based on the depth of *lcs* and the taxonomy, or the gloss overlap in WordNet. Riensche *et al.* (2007) also adopted a similar strategy. Sheng *et al.* (2010) linearly combined the distributional similarity of two words using a PMI-based method with the sematic similarity score calculated using Wu and Palmer's (1994) method. Similarly, Gracia and Mena (2008) linearly combined the score calculated using the Cilibrasi and Vitanyi's (2007) measure with one that calculated using an overlap-based method.

Other hybrid methods employ scores given by individual methods as an integral part of the hybrid model, usually as some kind of features. To distinguish them from the above combination methods, we call this type of hybrid method the *integration* methods. Rodríguez and Egenhofer (2003) combined a gloss-based method with depth as a normalization factor. Given the gloss created for two concepts, semantic relatedness is determined by the common characteristics (set overlap) and non-common characteristics (set difference), calculated using an adapted Tversky (1977) set-based algorithm. Let $U(c_i, c_j) = gloss(c_i) \bigcap gloss(c_j)$ and $R(c_i, c_j) = gloss(c_i)/gloss(c_j)$, it is calculated as follows:

$$SemRel(c_1, c_2) = \frac{|U(c_1,c_2)|}{|U(c_1,c_2)|+\alpha(c_1,c_2)\cdot|R(c_1,c_2)|+(1-\alpha(c_1,c_2))\cdot|R(c_2,c_1)|} \tag{31}$$

and $\alpha(c_1, c_2)$ is the *relative importance factor* for non-common characteristics, calculated using the depth of two concepts in the taxonomy:

$$\alpha(c_1, c_2) = \begin{cases} \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)} & \text{if } depth(c_1) \le depth(c_2) \\ 1 - \dfrac{depth(c_1)}{depth(c_1) + depth(c_2)} & \text{otherwise} \end{cases} \tag{32}$$

This measure is asymmetric. The effect of weighting non-common characteristics with respect to the depth of two concepts is that relatedness from deeper concepts to shallower concepts is higher than the opposite, which is consistent with the

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

common perception of asymmetric semantic relatedness as discussed in Section 2. This method is later adapted by Petrakis and Varelas (2006) with different definitions of gloss.

Othman, Deris and Illias (2007) defined semantic similarity of two concepts in a taxonomy as a function of their individual distance to their *lcs*. This distance is calculated using a measure that combines the *IC*, depth, and local density in order to address specificity of concepts:

$$dist(c_1, c_2) = \sum_{i}^{l} D(c'_i) \cdot E(c'_i) \cdot (IC(c'_{i+1}) - IC(c'_i))) \tag{33}$$

where $c'_1, c'_2, ..., c'_l$ are the list of concepts along the path from c_1 to c_2, $D(c)$ and $E(c)$ are functions that return the depth and local density of the concept node, respectively. They are slightly modified based on the classic definition of depth and density. The distance is effectively the sum of weighted edges along the path from c_1 to c_2, which has a similar notion to *semantic path elaboration* proposed by Tsatsaronis *et al.* (2010). The final semantic similarity is computed as

$$SemSim(c_1, c_2) = 1 - min\left\{1, \frac{dist(c_1, lcs(c_1, c_2)) + dist(c_2, lcs(c_1, c_2))}{max_{c \in T}\{IC(c)\}}\right\} \tag{34}$$

Pozo *et al.* (2008) proposed to derive a taxonomic structure of terms from a corpus and apply path-based methods to the taxonomy. Their motivation is that even if a well-curated knowledge base is available, many structural relations may not be encoded but may be 'hidden' in a large corpus. Thus, they proposed to uncover hidden structural relations between terms by applying hierarchical clustering to the terms based on their distributional features observed from corpora. The approach consists of four steps: (1) representing each term extracted from a corpus by a vector of its contextual words, and compute pairwise similarity of terms using the cosine similarity function; (2) creating a connected graph of terms, where edges are established for two terms if the pairwise similarity is above a threshold; (3) apply spectral clustering to partition the graph; and (4) apply agglomerative hierarchical clustering to generate a clustering tree of terms, which is used as a taxonomy. Then the semantic distance between any terms is simply the depth of their *lcs*.

Han and Zhao (2010) proposed the *Structural Semantic Relatedness* (SSR) method, which makes use of three methods, including that by Lin (1998b), Milne and Witten (2008) and the NGD method by Cilibrasi and Vitanyi (2007). Given a collection of words, the method begins with a pre-processing step that calculates semantic relatedness for each pair of words using all three methods. Because of the coverage of background information resource used in each method, a pair may receive between one and three relatedness scores. For multiple scores, the score returned by the most 'reliable' method is used. This is arbitrarily determined in the order of preference as Lin (1998b), Milne and Witten (2006), and NGD. Next, the words are plotted in a connected graph where the score is used to weigh edges between them. Let $e(w_1, w_2)$ denote the weight of the edge connecting two words, $neighbor(w)$ returns the immediate neighbors (connected by an edge) of a word, the SSR is then

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

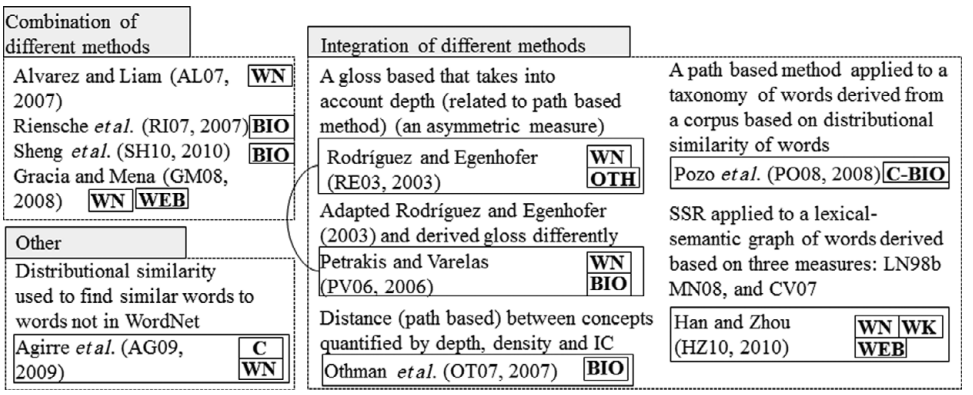


Fig. 11. A summary of hybrid methods and the initial background information resources used. WN: WordNet, WK: Wikipedia, BIO: biomedical knowledge bases, OTH: other structured knowledge bases, WEB: the web, C: a general-purpose corpus, C-BIO: a biomedical corpus.

computed as:

$$SemRel(w_1, w_2) = \alpha \cdot \sum_{w' \in neighbor(w_1)} \left(\frac{e(w_1, w')}{d_{w_1}} \cdot SemRel(w', w_2) \right) + \beta \cdot e(w_1, w_2) \quad (35)$$

where d_w is the degree of the node representing a word on the graph, and α and β are control parameters. The computation of SSR is recursive. It formulates the intuition that two words are similar if they are semantically related to a similar set of neighbors. The recursive equation can be solved by applying matrix algebra. The authors claimed that this method takes into account both explicit semantic relations and implicit semantic connections, and overcomes the coverage limitations of individual background information sources.

Another hybrid method that does not fall under either of the two categories is Agirre *et al.* (2009). They proposed to combine a WordNet-based method adapted from Hughes and Ramage (2007) with a distributional similarity method based on context similarity. The method adapted from Hughes and Ramage (2007) exploits the link structure of WordNet. However, because of the limited coverage of WordNet, a small proportion of testing data is not covered. To cope with this, for each word that is unknown to WordNet, they firstly applied the distributional similarity method to find several similar words. These are then used to substitute the unknown word and the WordNet-based approach is reapplied.

A summary of hybrid methods is shown in Figure 11.

4.7 Summary

In the previous sections we have discussed methods for computing lexical semantic relatedness in six different categories. The majority of these methods have assumed relatedness to be symmetric, although a number of asymmetric relatedness methods have also been proposed (Rodríguez and Egenhofer 2003; Petrakis *et al.* 2006; Schickel-Zuber and Faltings 2007). While new methods are constantly introduced, recent studies in this area can be characterized by several trends.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Categories. In terms of different categories, there is a clear preference for vector-based methods. For example, all of the vector-based methods are introduced from 2006 onwards, and they account for a large proportion (nearly 40%) of all methods introduced since then. This is consistent with the findings by Zesch and Guryvech (2010a). Hybrid methods have also become widely used.

Background information resources. Recent studies have explored a variety of new background information resources. Collaborative knowledge resources such as Wikipedia have been studied extensively, particularly by vector-based methods (Gabrilovich and Markovitch 2007; Milne and Witten 2008; Zesch *et al.* 2008b; Yeh *et al.* 2009). Studies have also been carried out to adapt previous methods based on WordNet to these resources (Strube and Ponzetto 2006; Zesch and Gurevych 2010a). Distributional similarity methods have exploited the Web as an alternative background corpus to classic pre-compiled corpora (Chen *et al.* 2006; Matsuo *et al.* 2006; Cilibrasi and Vitanyi 2007). In addition, new trends such as combining different knowledge bases to complement each other (Han and Zhao 2010; Zhang *et al.* 2011) and incorporation of temporal usage pattern of concepts (Radinsky *et al.* 2011) have also been suggested.

Multilinguality. Methods of lexical semantic relatedness are primarily introduced for English. As discussed before, most semantic relatedness methods can be described in a generic way and adapted to different domains and languages. However, in practice, this has been rarely addressed. A few recent studies have explored this direction. Gurevych and Niederlich (2005) adapted the methods by Resnik (1995), Jiang and Conrath (1997), and Lin (1998b) to a German lexical knowledge base and tested them on a German dataset. Zesch *et al.* (2008b) and Zesch and Gurevych (2010a) also tested a few state-of-the-art methods for German. Hassan and Mihalcea (2009) tested their method in cross-lingual semantic relatedness, i.e., determining relatedness between a term from one language (e.g., *factory*) and a term from another language (e.g., *lavoratore* in Italian, *worker* in English). They exploited the inter-language links in Wikipedia to map terms in one language with another. The method was tested for Spanish, Romanian, Arabic, as well as English. Liu and Chen (2010) adapted the method by Wu and Palmer (1994) and the cosine similarity between concept vectors to compute semantic relatedness am Chinese-named entities using Wikipedia.

In Sections 5 and 6, we discuss the evaluation of semantic relatedness methods and lessons learnt.

5 Evaluation of lexical semantic relatedness methods

Methods for measuring lexical semantic relatedness are typically evaluated by two types of approaches: *in-vitro* and *in-vivo*. In *in-vitro* experiments semantic relatedness scores of concepts or words are compared directly against a gold standard. In *in-vivo* experiments the methods are evaluated indirectly by the performance of an application built on top of it.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

5.1 In-vitro *evaluation*

In-vitro evaluation can be done by mathematical analysis and correlation with human judgement. Lin (1998b) proposed a method to assess semantic relatedness methods against certain properties, e.g., whether the score between two concepts increases with their commonality and decreases with their difference. However, this does not assess how well the method performs on real data. The predominant *in-vitro* evaluations are based on correlation with human judgement, also known as the gold standard. Typically, a dataset containing a set of pairs of concepts or words are presented to human judges, who subjectively estimate the relatedness between each pair within a certain scale. Next, the semantic relatedness method is applied to the same dataset. The results are correlated against the human judgement to derive an indication of the accuracy of the method.

5.1.1 *Correlation measures*

Two correlation functions are often used in the literature, the Pearson correlation coefficient and the Spearman rank-order correlation coefficient. The Pearson correlation compares the scores computed by a semantic relatedness method with the numeric scores of the gold standard. For example, the pair *table-chair* might get a human judgement of 8/10. With the Pearson correlation, a machine-computed score of 7/10 will be awarded higher than a score of 4/10, since the first score is closer to the gold standard. The Pearson correlation returns a value of 1 for perfect correlation and a value of 0 for no correlation. Given X the list of scores assigned to a list of term pairs by humans, and Y the scores assigned to the same list by a semantic relatedness method, it is calculated as given below:

$$\text{Pearson}(X, Y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \sqrt{\sum (Y - \overline{Y})^2}} \tag{36}$$

The Spearman rank-order correlation coefficient is based on rankings of data series. For this method, the list of term pairs are ranked by their scores given by the human judge and a measure, respectively, and correlation strength is assessed based on how well the ranking given by the measure resembles that by the human. Effectively it can be calculated by the same formula for the Pearson correlation coefficient by simply replacing the lists of scores X and Y with the lists of ranks of these scores.

Zesch and Gurevych (2010a) discussed the limitations of each correlation measures and argued in favor of the Spearman measure. Firstly, the Pearson correlation function is very sensitive to outliers – a single outlier may produce a significantly different result. They further justified this in an experiment and showed that on the dataset by Miller and Charles (1991), the method by Lesk (1986) was significantly penalized by a single word pair that received an extraordinarily high score representing an outlier in the dataset. Secondly, it measures the strength of the linear relationship between two data series and can produce flawed results when the relationship between human judgements and computed scores is nonlinear.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Thirdly, it requires normal distribution of two random variables (scores given by a human judge and a measure) and scores to be normalized within certain interval scales. However, studies have shown that semantic relatedness scores are not always interval-scaled (Budanitsky and Hirst 2006; Zesch and Gurevych 2007).

In contrast, the Spearman correlation is a more robust measure and does not suffer from these limitations. However, it is known from the literature that it may yield skewed results on datasets with many tied ranks.

5.1.2 Gold standards

Gold standard datasets for correlation analysis are typically created by asking multiple human annotators to rate the semantic relatedness of a pair of concepts or terms within a certain scale, and then averaging their interpretations. Since the interpretation is subjective, inter-annotator agreement (IAA) should also be studied.

General domain English datasets. Since studies on semantic relatedness are predominantly focused on English, a variety of datasets have been created for evaluation. Among these, four datasets and their variants have been widely used. These are discussed in details in Zesch and Gurevych (2010a). Briefly, these contain the Rubenstein and Goodenough (1965) dataset of sixty-five pairs of nouns (RG65), the Miller and Charles (1991) dataset that is a subset of the RG65 dataset and contains thirty pairs (MC30), the Finkelstein *et al.* (2002) dataset containing 353 pairs of words (Fin353), and the Yang and Powers (2006) dataset containing 130 verb pairs (YP130). In addition, Gracia and Mena (2008) created a dataset of thirty pairs of common nouns. No IAA figures were reported for the original RG65 and MC30 datasets. Pirró and Seco (2008) and Resnik (1995) recreated these datasets with IAA analyses. Also, the Fin353 dataset contains two subsets, each containing 153 pairs and 200 pairs, respectively, and annotated by different groups of annotators. Zesch and Gurevych (2010a) carried out IAA analysis and discovered largely varying figures for the two subsets, and therefore suggested treating them separately (Fin153 and Fin200) in evaluation. The RG65, MC30, and YP130 are assessed based on similarity, and are therefore used originally for evaluating semantic similarity methods. For the same purpose, Agirre *et al.* (2009) created a semantic similarity dataset based on Fin353. This (AG203) contains 203 pairs of terms from Fin353 each rescored for their similarity rather than relatedness.

One widely known limitation of these datasets is that they do not contain multiword phrases, and only the Fin153 and Fin200 datasets contain very limited number of named entities. To address this issue, Ziegler *et al.* (2006) created two English datasets of concept instances and named entities. The first dataset (Zie25) contains twenty-five pairs, annotated by twenty-three people; the second (Zie30) contains thirty pairs, annotated by fifty-one people. Eighty-seven percent of annotators are German native speakers. These datasets have rarely been used. Radinsky *et al.* (2011) created a dataset of 280 pairs of terms (Rad280) containing both common English words and entities. It was then annotated manually by inviting participants through the Amazon Mechanic Turk service. Table 2 summarizes these datasets.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Table 2. *General domain English datasets used for evaluating lexical semantic relatedness methods based on correlation studies*

Dataset	Pairs	PoS	Similarity/Relatedness	IAA
RG65	65	N	Similarity	0.80
MC30	30	N	Similarity	0.90
YP130	150	V	Similarity	0.87
Fin153	153	N, V, A	Relatedness	0.73
Fin200	200	N, V, A	Relatedness	0.55
AG203	203	N, V, A	Similarity	–
Zie25	25	–	Relatedness	–
Zie30	30	–	Relatedness	–
GM30	30	N	Relatedness	–
Rad280	280	N, V, A	Relatedness	–

Table 3. *Biomedical English datasets used for evaluating lexical semantic relatedness methods based on correlation studies*

Dataset	Pairs	PoS	Similarity/Relatedness	IAA
MeSH-36	36	N	Relatedness	–
Ped-30	30	N	Similarity	0.85
Pak724	724	N	Both	0.24–0.63

General domain other languages. Compared with English, datasets for other languages are lacking. Gurevych (2005) translated the RG65 and MC30 datasets (Gur65) into German, and expanded them to a total of 350 word pairs. Zesch and Gurevych (2006) later created a new dataset of 222 German word pairs. Both datasets contain nouns, verbs, and adjectives. Cramer and Finthammer (2008) created a German dataset of 600 pairs of nouns. Hassan and Mihalcea (2009) translated the MC30 and Fin353 datasets into Spanish, Romanian, and Arabic. Liu and Chen (2010) created a dataset of twenty-eight pairs of named entities for Chinese. Note that translated datasets may be biased toward the original language (Cramer and Finthammer 2008), since in many cases the translation is not a one-to-one mapping.

Biomedical datasets. While the majority of general-purpose semantic relatedness methods have been evaluated using the above datasets, few of them have been evaluated in specialized domains, which is partly due to the lack of domain-specific datasets in similar forms. To our knowledge, in the biomedical domain, three datasets are available for evaluating semantic relatedness between terms, while two of them have been empirically used (Table 3). Petrakis *et al.* (2006) compiled a set of forty-nine MeSH term pairs and asked twelve experts to assess the relatedness of these pairs. Pairs with standard deviation above 0.8 were excluded, resulting in a total of thirty-six term pairs (MeSH36). Pedersen *et al.* (2007) created a set of 120 pairs of medical terms extracted from the SNOMED-CT terminology annotated by thirteen

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

experts. However, they obtained a very low IAA of 0.51 on this dataset. To create a more reliable test bed, they selected a subset of thirty pairs and asked another two groups of experts (three and nine annotators, respectively) to re-annotate them. The final dataset (Ped30) has IAA figures of 0.68, 0.78, respectively, and a cross-group IAA of 0.85. Pakhomov *et al.* (2010) performed experiments in which eight medical residents are required to annotate a set of 724 pairs of clinical terms (Pak724). The terms are extracted from UMLS and cover three semantic categories: disorders, symptoms, and drugs. Both intra-category and inter-category pairs are created. All datasets contain multi-word terms. Annotators are required to score pairs of terms by their relatedness and similarity separately, thus creating benchmarked datasets for both semantic relatedness and similarity methods. Due to time limit, 81% of the dataset was completed for the relatedness annotation, and 78% was completed for the similarity annotation. They found the tasks to be very challenging because only moderate IAA has been achieved. For the relatedness annotation task, an IAA between 0.45 and 0.56 (inter- and intra-categories) is reported; while for the similarity annotation task, this is between 0.24 and 0.63.

On the other hand, evaluation of semantic similarity methods in the biomedical domain is often based on a comparison against the result of a different approach such as gene sequence similarity or family similarity. Firstly, semantic similarity between gene products is calculated usually taking into account the semantic similarity between the terms used to annotate them in the corpora. The hypothesis is that gene products are similar if they share similar properties. A large number of methods are available for this task, which, however, is not the focus of this survey. Details of this may be found in the survey by Pesquita *et al.* (2009). Next, the sequence similarity and family similarity between gene products are measured using a different approach such as Altschul *et al.* (1997). The sequence or family similarity is then used as the gold standard against which the prediction based on semantic similarity is evaluated using a correlation metric. The hypothesis is that genes that share similar sequences or families should also be semantically similar.

5.1.3 Empirical results

Due to the large number of methods and various background information resources used, a comprehensive comparative evaluation can be very difficult. Firstly, only a very small portion of these methods – typically earlier (before 2000) WordNet-based methods – have been implemented and made publicly available. Re-implementation of other methods requires not only knowledge about the algorithms but also the background information resources used. Secondly, different methods that depend on the same background information resource may have used different versions, making the direct comparison difficult. The problem can be more acute for methods based on Wikipedia, which has been growing almost exponentially since its introduction. Zesch and Gurevych (2010b) have shown that the growth of Wikipedia can have a major impact on the coverage of a method, and smaller effect on accuracy.

Some studies (Curran and Moens 2002; Strube and Ponzetto 2006; Bollegala *et al.* 2007; Zesch *et al.* 2008b; Zesch and Gurevych 2010a) have addressed these issues

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

by re-implementing other methods and adapting them to a uniform background information resource. However, the majority of studies simply compare their results against others using previously published figures, ignoring the difference in terms of background information resources.

In this work we summarize previously reported evaluations based on the datasets introduced above. Some methods have been adapted to different background information resources and retested on the same datasets in later studies. In this case, we refer to the corresponding original work if it has been evaluated on the dataset; or if not, the most recent work that has included an evaluation on that dataset. For studies that evaluate different variants of a method (e.g., parameter configurations, learning algorithms), we present the highest figures obtained. We present a summary of evaluations on the general purpose datasets in Table 4, where each method is indexed by an acronym originally introduced in Figures 2–4, 8, 10, and 11. A score calculated using the Pearson measure is denoted by r, and a score calculated using the Spearman measure is denoted by ρ.

In addition, a small number of methods have been evaluated using biomedical datasets in Table 3. The majority of these (except Zhang *et al.* 2011) have been built on or adapted to specific biomedical background information resources. Table 5 shows the evaluation of these methods and the used corresponding background information sources.

5.1.4 Interpretation of results

Datasets and domains. As shown in Tables 4 and 5, the majority of methods have been evaluated in the general domain, while only a few methods have been evaluated on biomedical datasets. On the one hand, only a very small fraction of generic methods – typically earlier WordNet-based methods – have been adapted to biomedical resources and evaluated on relevant datasets; on the other hand, a large proportion of methods developed in the biomedical domain have only been evaluated using domain-specific approaches, making it difficult for comparative studies.

For evaluation in the general domain, a large number of methods have not been assessed using the gold standard datasets. As we shall discuss in Section 5.2, many of these have been evaluated indirectly in certain tasks built upon the semantic relatedness method. The use of different datasets for evaluation is very imbalanced. The MC30 and RG65 datasets have been used by the majority of methods, particularly the path-based and IC-based methods. Despite the chronological order of the publication of relevant work and datasets, this can be due to the fact that most of these methods address specifically semantic similarity, which the two datasets are tailored for. In contrast, Fin353 and its subsets have been used largely for evaluating vector-based methods, which address the more general relatedness. Other datasets, including AG203, GM30, Rad280, Zie25, and Zie30 have been rarely used.

Effects of correlation measures. In terms of the choice of correlation measures, where both are used for a method (RA89, WP94, HS98, LC98, TS10, RE95, JC97, LN98b, LE86, PP06, GM07, HM09, and SZ11), we can see that in general the difference

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Table 4. *Evaluation in the general domain*

Method	Dataset							SoF
	MC30	RG65	YP130	Fin153	Fin200	Fin353	Other	
Path-based methods								
RA89	.76 *r*	.79 *r*	.74 *r*	**.38** *r*	**.36** *r*	–	–	[1]
	.75 *ρ*	.79*ρ*	.64 *ρ*	.33 *ρ*	.24 *ρ*	–	–	
WP94	.78 *r*	.80 *r*	.76 *r*	.28 *r*	.24 *r*	–	–	[1]
	.77 *ρ*	.78 *ρ*	.67 *ρ*	**.38** *ρ*	.28 *ρ*	–	–	
HS98	.67 *r*	.73 *r*	.70 *r*	.35 *r*	.35 *r*	–	–	[1]
	.76 *ρ*	.79 *ρ*	.61 *ρ*	.33 *ρ*	.32 *ρ*	–	–	
LC98	.76 *r*	.84 *r*	.74 *r*	.34 *r*	.31 *r*	–	–	[1]
	.75 *ρ*	.79 *ρ*	.64 *ρ*	.33 *ρ*	.24 *ρ*	–	–	
JS03	.88 *r*	.82 *r*	–	–	–	–	–	o.w.
LI03	.89 *r*	–	–	–	–	–	–	o.w.
YP05	.92 *r*	**.90** *r*	**.83** *r*	–	–	–	–	o.w., [2]
LI07	**.93** *r*	.89 *r*	–	–	–	–	–	o.w.
SF07	.91 *r*	–	–	–	–	–	–	o.w.
TS10	.86 *r*	.86 *r*	–	–	–	–	–	o.w.
	.86 *ρ*	.88 *ρ*	–	–	–	–	–	
WH11	.81 *r*	.82 *r*	–	–	–	–	–	o.w.
IC-based methods								
RE95	.79 *r*	.81 *r*	.69 *r*	**.36** *r*	**.31** *r*	–	–	o.w., [1]
	.72 *ρ*	.74 *ρ*	.61 *ρ*	.35 *ρ*	.26 *ρ*	–	–	
JC97	.83 *r*	.71 *r*	.61 *r*	.34 *r*	.28 *r*	–	–	o.w., [1]
	–	.58 *ρ*	.68 *ρ*	.28 *ρ*	.10 *ρ*	–	–	
LN98b	.83 *r*	.72 *r*	**.76** *r*	.31 *r*	.27 *r*	–	–	o.w., [1]
	–	.62 *ρ*	.67 *ρ*	.27 *ρ*	.17 *ρ*	–	–	
SE04	**.84** *r*	–	–	–	–	–	–	o.w.
PI09	–	**.91** *r*	–	–	–	–	–	o.w.
Gloss-based methods								
LE86	.36 *r*	.34 *r*	.39 *r*	.20 *r*	.28 *r*	–	–	[1]
	.78 *ρ*	.72 *ρ*	.64 *ρ*	.47 *ρ*	.33 *ρ*	–	–	

between the two metrics is nonsignificant, except in the case of LE86. As discussed before, according to Zesch and Gurevych (2010a) the low *r* value is caused by a single word pair that gives an extraordinarily high score in the dataset, confirming their argument about its sensitivity to outliers. However, the effect of tied ranks on the Spearman correlation has not been investigated.

Accuracy. In terms of the general domain, the overall picture is rather inconclusive. In general, there is no single method that always outperforms others on any dataset. For each of the most frequently used datasets, the best performing methods are LI07

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Table 4. *(Continued)*

Method	Dataset							SoF
	MC30	RG65	YP130	Fin153	Fin200	Fin353	Other	
Vector-based methods								
ZI06	–	–	–	–	–	–	Zie30 .75 *r*	o.w.
PP06	**.91** ρ	**.90** ρ	.38 *r*	.11 *r*	.09 *r*	–	–	o.w.,[1]
	–	–	.39 ρ	.10 ρ	.10 ρ	–	–	
GM07	.72 ρ	.82 ρ	.30 *r*	.50 *r*	.29 *r*	.75 ρ	–	o.w.,[1]
	–	–	.29 ρ	.61 ρ	.28 ρ	–	–	
HR07	.90 ρ	.82 ρ	–	–	–	–	–	o.w.
ZE08b	.84 ρ	.84 ρ	.65 ρ	.70 ρ	**.60** ρ	–	–	o.w.
MW08	.70 ρ	.64 ρ	–	–	–	.69 ρ	–	o.w.
HM09	.75 ρ	–	–	–	–	.71 ρ	–	o.w.
	.58 *r*	–	–	–	–	.55 *r*	–	
YE09	.82 ρ	–	–	–	–	.63 ρ	–	o.w.
YP10	–	–	–	–	–	.71 ρ	–	o.w.
	–	–	–	–	–	.68 *r*	–	
GO10	–	–	–	–	–	.70 ρ	–	o.w.
HA10	–	.86 ρ	–	–	–	.62 ρ	–	o.w.
WP11	–	–	–	–	–	.50 ρ	–	o.w.
RA11	–	–	–	–	–	**.80** ρ	Rad280 .63 ρ	o.w.
ZH11	.80 ρ	.74 ρ	–	**.75** ρ	.54 ρ	–	–	o.w.
SZ11	–	.86 ρ	.70 ρ	–	–	–	–	o.w.
	–	**.90** *r*	**.72** *r*	–	–	–	–	
Distributional similarity methods								
SH06	.58 *r*	–	–	–	–	–	–	[3]
CH06	.69 *r*	–	–	–	–	–	–	[3]
BO07	**.81** *r*	–	–	–	–	–	–	[3]
Hybrid methods								
RE03	.71 *r*	–	–	–	–	–	–	[4]
PV06	**.74** *r*	–	–	–	–	–	–	[4]
AL07	.90 *r*	.91 *r*	–	–	–	–	–	o.w.
GM08	–	–	–	–	–	–	GM30 .74 ρ	o.w.
AG09	–	**.96** ρ	–	–	–	**.78** ρ	AG203 .83 ρ	o.w.

Sources of figures (SoF) – o.w.: original work; [1] Zesch and Gurevych (2010a); [2] Yang and Powers (2006); [3] Bollegala et al. (2007); and [4] Petrakis et al. (2006). For each category of methods, on each dataset and with each correlation metric, the best performing method is highlighted in bold.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Table 5. *Evaluation based on biomedical datasets*

Method	Category	Ped30	MeSH36	BG Info. Resource	S. of F.
AN06	Path	.86 *r*	.83 *r*	MeSH	[1]
RA89	Path	.74 *r*	.77 *r*	MeSH	[1]
WP94	Path	.79 *r*	.84 *r*	MeSH	[1]
LC98	Path	.86 *r*	.82 *r*	MeSH	[1]
LI03	Path	.85 *r*	.83 *r*	MeSH	[1]
RE95	IC	–	.72 *r*	MeSH	[2]
JC97	IC	–	.71 *r*	MeSH	[2]
LN98b	IC	–	.72 *r*	MeSH	[2]
PI09	IC	–	.73 *r*	MeSH	[2]
BA11	Gloss	.73 *r*	–	SNOMED-CT	o.w.
ZH11	Vector	.67 ρ	.75 ρ	WordNet + Wikipedia	o.w.
RE03	Hybrid	–	.71 *r*	MeSH	o.w.
PV06	Hybrid	–	.71 *r*	MeSH	o.w.

S. of F.: Source of figures; BG info.: background information; [1] Al-Mubaid and Nguyen (2006), [2] Pirró (2009).

(path-based) on MC30, AG09 (hybrid, distributional similarity combines vector-based) on RG65, YP05 (path-based) on YP130, RA11 (vector-based) on Fin353, ZH11 (vector-based) on Fin153, and ZE08b (vector-based) on Fin200. Within each category of methods, the modifications introduced in later methods generally lead to higher accuracies than the earlier basic methods. For example, in the case of path-based methods, those (WP94, LC98, LI03, LI07, SF07, TS10, WH11) addressing concept specificity in a hierarchy generally obtain higher accuracies than those (RA89, HS98) ignoring specificity; and in the case of IC-based methods, later models (JC97, LN98b, SE04, and PI09) have further improved over their ancestor (RE95). In particular, Pirró (2009) demonstrated the superiority of the intrinsic IC (Seco *et al.* 2004) to the original definition by Resnik (1995) by comparative experiments. In terms of performance per dataset, most path-based methods obtain good results on the MC30, RG65 and YP130 datasets, but rather poor results on the Fin353 dataset and its subsets. This can be understood taking into account that these methods are generally designed for measuring semantic similarity rather than relatedness. For the same reason, similar observations are also found in IC- and gloss-based methods. In contrast, performance of vector-based methods is more balanced across all datasets (except PP06). They obtain comparable accuracies on the similarity datasets, and much higher accuracies on the relatedness datasets. Taking these observations into consideration, we argue that vector-based methods are generally superior to other categories in this task.

In terms of the biomedical domain, it is difficult to compare by categories due to lack of statistics. However, the available figures tend to suggest that path-based methods have generally higher accuracies than others. This is possibly due to the high quality of knowledge bases available in the biomedical domain.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Effect of background information resources. Since semantic relatedness methods depend on background information resources, the quality of the chosen resource will have a major impact on the accuracy of the method. The choice of background information resources and their strengths and weakness will be discussed in Section 6; here we discuss their possible effects on accuracies using statistics in Tables 4 and 5.

In the general domain, the most extensively used background information resources are WordNet and Wikipedia. Considering the figures in Table 4 and the underlying background information resources of each method shown in Figures 2–4, 8, 10, and 11, there is the tendency that suggests methods based on WordNet generally score better on the MC30, RG65 and YP130 datasets, while methods based on Wikipedia score better on the Fin353 and its subset datasets. This is generally consistent with the findings by Strube and Ponzetto (2006), who carried out comparative studies by adapting several WordNet-based methods to Wikipedia. This possibly suggests that different focuses of knowledge in WordNet and Wikipedia have biased them toward measuring similarity and general relatedness. To a certain extent, this hypothesis is supported by a later work by Ponzetto and Strube (2011), in which they re-adapted several WordNet-based methods to a finely induced taxonomy based on the Wikipedia category tree. They showed that by cleaning, classifying the links in Wikipedia category tree, and focusing on the *IS-A* relation, the adapted methods outperformed their WordNet originals. Zesch and Gurevych (2010a) also carried out comparative studies based on WordNet and Wikipedia. However, they reached a rather contradictory conclusion that the differences caused by the two knowledge bases are rather insignificant. Although WordNet gives better accuracies for verb pairs, Wikipedia is also strongly competitive.

Another observation is the inconsistency of results reported by some studies for the same method. For example, JC97 and LN98b reported a correlation of 0.87 and 0.83 on MC30 respectively in their original work; however, Zesch and Gurevych (2010a) have reported much lower figures (0.68 and 0.67) for the same method on the same dataset. Similarly, the author reported 0.61 ρ and 0.28 ρ for a re-implementation of the GM07 method (Gabrilovich and Markovitch 2007) on the Fin200 and Fin153 datasets, which seem to contradict the 0.75 ρ correlation on the Fin353 dataset reported in the original work. A later study by Bär, Zesch and Gurevych (2011) suggested that this could be the effect of pruning feature vectors for specific datasets in the original work, which may have caused overfitting. Meanwhile, another potential cause could be the use of different versions of knowledge bases in these studies.

For the biomedical domain, semantic relatedness is predominantly computed based on domain-specific resources. One exception is the ZH11 method (Zhang *et al.* 2011), which produces comparable results by harnessing only several general-domain knowledge bases. This opens the possibility of porting generic methods to domain-specific tasks where domain-specific resources are unavailable. In addition, a comparative evaluation by Al-Mubaid and Nguyen (2006) has shown that much higher accuracies can be obtained using MeSH than using SNOMED-CT for the same method. Their experiments using five-path-based methods showed that the

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

accuracies dropped up to 0.50 and 0.23 on the Ped30 and MeSH36 datasets, respectively, when the underlying knowledge base was changed from MeSH to SNOMED-CT. According to the authors this is attributed to the higher level of specificity (granularity) in the SNOMED-CT concept hierarchy, which has penalized methods that do not address specificity adequately.

5.2 In-vivo *evaluation*

In addition to *in-vitro* evaluation based on the gold standard datasets, a large number of studies carry out *in-vivo* evaluation by assessing the performance of an application built on top of lexical semantic relatedness methods. Due to various choices and diverse datasets used, a direct comparison is infeasible. Instead, we briefly introduce relevant work and frequently use datasets as general pointers for readers.

Methods of lexical semantic relatedness can be adapted to measure *text similarity*, i.e., similarity between long text passages such as sentences and paragraphs. It is often considered a task highly related to semantic relatedness but essentially different. At sentence level, similarity concerns semantic, part-of-speech, and syntactic information (Li *et al.* 2009). At paragraph or document levels, Bär *et al.* (2011) argued that similarity should be defined along three dimensions: structure, e.g., the order of sections; style, e.g., grammar, usage; and content, i.e., facts mentioned in the texts and their relations. Only the content dimension concerns the semantic level of similarity. In general, text similarity addresses a broader sense of similarity than purely semantic. The semantic level of text similarity can be addressed by adapting methods of lexical semantic relatedness, in which case we consider text similarity as an application of lexical semantic relatedness. Often, semantic relatedness between the component terms of texts are computed and then aggregated to derive a score as text similarity (Mihalcea, Corley and Strapparava 2006). Alternatively for long texts, it can simply be evaluated by comparing the vocabularies (Lee, Pincombe and Welsh 2005).

A number of datasets have been created for evaluating text similarity measures (in the semantic sense). Li *et al.* (2006) introduced sixty-five sentence pairs based on the noun pairs created by Rubenstein and Goodenough (1965). Each noun was replaced by its definition in a reference English dictionary. Mohler and Mihalcea (2009) created a dataset based on computer science assignments. It contains twenty-one short questions with a reference answer and also 630 student answers for all questions. The answers were graded based on the extent to which the student answers matched the reference answer. Lee *et al.* (2005) created a collection of 1,225 pairs of broadcast news articles and conducted experiments to assign semantic relatedness scores to each of them based on human judgements. The dataset has been used by, for example, Gabrilovich and Markovitch (2007) for evaluating their ESA method, and Gouws *et al.* (2010). The Microsoft Paraphrase Corpus (Dolan, Quirk and Brockett 2004) is a collection of 1,725 pairs of statements; each assigned a semantic relatedness score determined by human judges. It has been used for evaluating the semantic relatedness methods of Yazdani and Popescu-Belis (2010).

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Another popular application is solving *word choice questions*. A typical question presents a target word and four choices, from which the most similar must be selected. This type of question is often found in many English language tests such as Test of English as a Foreign Language (TOEFL). It has been adopted by a number of studies, including Jarmasz and Szpakowicz (2003), Ruiz-Casado *et al.* (2005), Tsatsaronis *et al.* (2010), and Zesch and Gurevych (2010a). Maguitman *et al.* (2005) tested their method in a similar form. However, instead of selecting the most similar term from candidates, their task is selecting the most similar webpage (typically about specific entities) from candidate pages.

Word Sense Disambiguation (WSD) concerns associating words to their underlying meanings based on their contexts. Semantic relatedness methods have been applied in this task where the hypothesis is that the desired sense of a polysemous word is the one that is highly related to those of other words in its context. The SENSEVAL[8] public evaluation publishes benchmarking datasets for sense disambiguation of words. In 2007, it was renamed as SEMEVAL. Many studies, including Mihalcea and Moldovan (1999), Banerjee and Pedersen (2003), Patwardhan and Pedersen (2006), and Gracia and Mena (2008), have evaluated their methods using the WSD datasets published by SENSEVAL. The method by Lesk (1986) and one based on Rada *et al.* (1989) were evaluated in a WSD task by Navigli (2006). Some methods have been evaluated by *Named Entity Disambiguation* (NED), where an ambiguous name mention (proper nouns) must be associated with one specific entity. Cucerzan (2007) and Turdakov and Velikov (2008) each created a dataset using Wikipedia. Han and Zhao (2010) evaluated their method using the person name disambiguation datasets created by the SEMEVAL competition. A comprehensive survey of WSD and NED methods and relevant datasets can be found in Navigli (2009). Many semantic relatedness-based methods are included and details of the SENSEVAL and SEMEVAL evaluation datasets are provided.

Sense clustering is also a common technique to evaluate semantic relatedness methods. Usually semantic relatedness between words is calculated to create a similarity matrix, to which a clustering algorithm is applied to split data into clusters based on the scores. In the general domain, Matsuo *et al.* (2006) and Bollegala *et al.* (2007) created gold standard clusters of named entities based on the categories in the Open Directory Project.[9] Cilibrasi and Vitanyi (2007) evaluated their method by hierarchical clustering of a manually created dataset of words. In the biomedical domain, Speer *et al.* (2004) and Wang *et al.* (2007) clustered gene products based on semantic relatedness between GO terms that are used to annotate the functions related to each gene product.

In *Information Retrieval* (IR), semantic relatedness can be used as a ranking method to improve search accuracy. It was first used as a task to evaluate semantic relatedness methods in Rada *et al.* (1989). The method is used to calculate relatedness between queries and documents, the scores of which are then used to rank documents. Rada *et al.* (1989) used the PubMed corpus and human

[8] http://www.senseval.org/, last retrieved on 16 March 2012.
[9] http://www.dmoz.org/, last retrieved on 16 March 2012.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

judges in the experiment. Lord *et al.* (2003b) built a tool for searching similar proteins in a gene annotation database. Rodríguez and Egenhofer (2003) evaluated their method by searching for similar concepts and/or entities in an ontology. Schickel-Zuber and Faltings (2007) tested their method in a movie search task, in which the semantic relatedness method was used to rank results to the best match against users' preferences. Sahami and Heilman (2006) applied semantic relatedness to query expansion, a technique for improving search accuracy. Egozi, Markovitch and Gabrilovich (2011) applied the ESA (Gabrilovich and Markovitch 2007) method to both document indexing and query expansion. ESA was used to find concepts related to queries and text documents. The concepts are then used to enrich the representations of queries and documents created based on the traditional bag-of-words model. Tested with standard IR approaches, the enriched representations have led to significant improvement over the traditional bag-of-words model.

Coreference resolution is the task of identifying expressions that refer to same entities. For example, in the sentences, *Google is a multinational IT corporation. The company was originally founded by Larry Page and Sergey Brin*, *Google* and *The company* are coreferent expressions, since they refer to the same entity. Semantic relatedness can be used for this task because the expressions used to refer to the same entity should have a certain semantic relation (Yang and Su 2007). Ponzetto and Strube (2006) used semantic relatedness measures to compute a score for each pair of referring expressions. The score is used as a feature in a supervised classification model for coreference resolution. They tested a list of semantic relatedness methods described in Strube and Ponzetto (2006) for this task. Yang and Su (2007) also used semantic relatedness scores as features but computed relatedness between two expressions using the PMI measure. Lee *et al.* (2011) exploited the lexical chains defined for two terms in WordNet, based on a similar rationale as that of path-based semantic relatedness methods. They encoded the hypernymy and synonymy relations between synsets of terms as features in a supervised classifier. Coreference resolution is a well-established research field and standard evaluation datasets have been created. Among these, the most often used include the datasets released in the Message Understanding Conference 6 (MUC6) and MUC7 (Chinchor 2001; Chinchor and Sundheim 2003), and the dataset by Mitchell *et al.* (2003).

Semantic similarity methods have also been tested in tasks related to *ontology construction and matching*. In most cases, the semantic similarity between a question term or concept and existing terms or concepts in a reference ontology are calculated, based on which it is linked to the reference ontology. Pekar and Staab (2002) evaluated their method based on its application in classifying terms into a domain-specific taxonomy GETESS (Staab *et al.* 1999). Lin (1998a) and Curran and Moens (2002) applied their methods to learning thesauri from a set of terms. Gracia and Mena (2008) tested their method in an ontology matching task, where terms in an ontology are matched with those in another based on semantic similarity between them.

Malapropism detection is an application of semantic relatedness to detect misused and misspelled words in text passages. For example, *word* in the sentence *She plans*

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

to spend her gap year travelling around the word is a misused word. It exploits the assumption that naturally coherent text passage should contain instances of related pairs of words (Budanitsky and Hirst 2006). The datasets for this type of experiments are usually based on that by Hirst and St-Onge (1998), who collected 500 articles from WSJ and artificially replaced every 200th word in the collection with a malapropism.

In addition, evaluation in the biomedical domain often adopts some *domain-specific* applications. In addition to the approach based on gene sequence or family similarity (Couto *et al.* 2005; Lei and Dai 2006; Schlicker *et al.* 2006; Pozo *et al.* 2008), some evaluate semantic relatedness methods by predicting gene interactions based on their functional similarity and predicting gene functions (Ye *et al.* 2005; Yu *et al.* 2005; Schlicker *et al.* 2006; Wu *et al.* 2006). Since these experiments are non-generalizable across domains, this paper does not discuss the details. Instead, a comprehensive summary of the evaluation in the biomedical domain can be found in Pesquita *et al.* (2009).

6 Discussion

In previous sections we have surveyed lexical semantic relatedness methods in both the general and biomedical domains and discussed their rationales. We have described evaluation methodologies, compared the performance of different methods based on the published resources, and discussed the results. One remaining objective of this work is to enable researchers and practitioners to make the right choice of lexical semantic relatedness method in relevant tasks. This decision cannot be straightforward and will depend on many factors. In the next sections, we discuss such considerations from several perspectives: the limitations of different categories of methods, the availability of background information resources and the domains in question, and the purpose of the task. As we shall see, these factors are often not independent and must be considered jointly with others.

6.1 Limitations of different categories of methods

Path-based methods can be very sensitive to the taxonomic structure of a knowledge base and the density and depth of the structure. They 'heavily depend on the degree of completeness, homogeneity, and coverage of the semantic links' represented in the structure (Batet *et al.* 2011). As an example, in Section 5.1.3 we have shown that for the same set of path-based methods, changing the underlying knowledge base (MeSH) to one of a higher level of specificity (SNOMED-CT) can have a major impact on their accuracy. Strube and Ponzetto's (2006) study has shown similar observations when adapting path-based methods from the WordNet structure to a less strict hierarchical structure, the Wikipedia category tree. Earlier path-based methods have assumed uniform distance of any edges regardless of the specificity of relevant concepts. This has been the major issue since the assumption is found to be untrue in most real taxonomies. It has been the focus of research and addressed

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

in different ways in later methods. One remaining issue is that most path-based methods are based on a single path following a single type of relation, typically *IS-A*. As a result, other useful semantic evidences are overlooked (Wang *et al.* 2007) and such background information may be insufficient to represent conceptual distance or relatedness between concepts in a semantic network (Lee *et al.* 1993). It has been shown (Resnik 1995; Richardson and Smeaton 1995) that when compared with some of the other categories of methods, path-based methods can lead to spurious results, which may be partially attributed to this issue. Nevertheless, compared with other methods, path-based methods are generally simple (Budanitsky and Hirst 2006; Pedersen *et al.* 2007; Pirró 2009).

IC-based methods inherit some limitations of the path-based methods, since they exploit the ancestors of two concepts as background information, which is dependent on the taxonomic structure. As a result, they can be also sensitive to the chosen taxonomic structure and drop of accuracies of some IC-based methods has also been observed when they are adapted from WordNet to Wikipedia (Strube and Ponzetto 2006). To some extent, this sensitivity may be offset by the use of corpus statistics in some IC-based methods. The use of a large corpus provides additional useful background information and may have contributed to superior performance in some IC-based methods as shown in Batet *et al.* (2011). However, in some cases such corpora are not always available, especially in the clinic domain where patient records are often highly confidential. In this case, as well as to eliminate corpus pre-processing, methods that approximate the IC of a concept based on the taxonomic structure (e.g., intrinsic IC) may be preferred. Similar to path-based methods, IC-based methods ignore other potentially useful semantic evidence but employ only the *IS-A* relation. In addition, path-based and IC-based methods are generally more suitable for measuring semantic similarity because of their emphasis on hierarchical relations.

Gloss-based methods present arguably a cheaper alternative to other kinds of methods, since the gathering of background information and the computation are generally less intensive. Due to the lack of comparable evaluations, it is difficult to discuss their limitations with respect to their performance. However, the study by Zesch and Gurevych (2010a) has shown that the Lesk's method (Lesk 1986) can be equally sensitive to the underlying background information source as path-based and IC-based methods.

Vector-based methods are generally more extensible than others and adaptation across different background information resources are generally straightforward. As we have shown, most vector-based methods differ in terms of how the concept vectors are created while sharing a large degree of commonality in terms of algorithmic calculation. Zesch *et al.* (2008b) also showed that a concept vector-based method can be easily adapted across three different knowledge bases (WordNet, Wikipedia, and Wiktionary) and obtaining comparable results. In terms of performance, we have discussed that vector-based methods are more balanced for both semantic similarity and relatedness tasks. They are also less sensitive to underlying background information resources, as shown by the results in Zesch *et al.* (2008b) and Zesch and Gurevych (2010a). This may suggest that vector-based methods can be a

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

better option when addressing new domains and datasets. However, the methods of constructing concept vectors differ significantly, and may lead to substantial differences in their performances. Methods that construct concept vectors using other relevant concepts in the semantic graph (Gabrilovich and Markovitch 2007; Harrington 2010) require extensive pre-processing of the entire background information resource.

Distributional similarity methods offer a major advantage over other methods – the flexibility in the choice of background information resource. Semantic relatedness methods typically employ a structured knowledge base, whose structure and the coverage and completeness of knowledge may limit the capability of the methods. Distributional similarity, however, is in theory free from such limitation, since the underlying corpora can be substituted without incurring changes to the method. This also makes it easily extensible to other domains. Pre-processing a large corpus of millions of documents will be a major issue to be considered with these methods, as it creates substantial computational cost (Pantel *et al.* 2009). In addition, the intrinsic difference between distributional similarity and semantic relatedness should also be considered when they are used as a proxy for semantic relatedness.

Hybrid methods are usually created to combine the strength of different measures. Based on the results that are somewhat inadequate to draw a final conclusion, it seems that they can lead to marginal improvement to their purebred competitors. Thus, before committing to a hybrid method one should consider whether the complexity of the methods can be justified in the particular context. Another consideration is that hybrid methods may inherit the limitations of purebred methods that are combined, which might explain the decreased performance of some methods (Rodríguez and Egenhofer 2003; Petrakis *et al.* 2006) when they are compared against some path- and IC-based methods.

6.2 Background information resources and domains

Since computing semantic relatedness depends on some background information resources, the choice of such resources will have a major impact on the performance of methods. Apparently the choice of background information resources is often bound by the methods. Nevertheless, when multiple choices are available, we should consider the types of information encoded, the focus as well as the coverage, and match these against the requirements of the task.

Generally compared to unstructured corpora, structured knowledge bases encode explicit semantic and lexical relations between concepts and entities, which are essential for assessing relatedness and similarity. As discussed in Section 3.1, WordNet and Wiktionary are knowledge bases of common words, focusing on nouns, verbs, adjectives, and adverbs. They have very limited coverage of proper nouns and specialized concepts. For this reason, they can be a good choice for tasks related to words, such as WSD. Also, the availability of well-defined hierarchical relations allows the knowledge bases to be tailored for semantic similarity other than just relatedness. In comparison, Wikipedia is a vast knowledge base of concepts and

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

entities, which makes it better suited for tasks such as NED. However, the coverage of word knowledge can be very limited, as shown in Zesch and Gurevych (2010a) where methods based on Wikipedia obtained poor performance on the YP130 verb datasets. Its broad coverage of a large number of topics also suggests that Wikipedia can be a better choice for domain-specific tasks. As an encyclopedic resource, Wikipedia focuses on covering fact-like information related to concepts and entities and present them as articles. Although Wikipedia articles are intensively hyperlinked, the semantic relations represented by such links are undefined. The category tree used to tag the articles is also a non-strict taxonomy. For this reason, it may be better suited for measuring semantic relatedness, while using Wikipedia for semantic similarity may require adaptation in order to obtain competitive results to WordNet-based methods, which has been shown in Ponzetto and Strube (2011).

Research in the biomedical domain generally prefers biomedical knowledge bases. They usually encode hierarchical relations between concepts, and have been used mostly for measuring semantic similarity. Compared with general-purpose knowledge bases, the distribution of knowledge is more uneven, resulting in different densities of regions and incomparable connections (Pozo *et al.* 2008). This is largely due to the nature of biomedical science, where knowledge is constantly updated, and the empirical knowledge of each concept is highly variable (Li *et al.* 2010). As a result, methods exploiting the hierarchical structure of such knowledge bases should take into account its unbalanced structure, for example, by addressing specificity of concepts in path-based methods.

A major limitation of using structured knowledge bases is that their scope and coverage can limit the capability of methods. Besides, such resources are usually expensive to maintain, and often unavailable in specific domains. In contrast, unstructured corpora are generally much easier to obtain, which offers the advantage of easier domain adaptation when knowledge bases are unavailable. Background information is often gathered in an implicit form, such as co-occurrence statistics primarily used for distributional similarity. As a result, important semantic evidences encoded in a structured knowledge base are often neglected. Although some methods (Pozo *et al.* 2008; Gouws *et al.* 2010; Harrington 2010; Wojtinnek and Pulman 2011) have proposed to mine hierarchical structures of words from corpora to address this, this often comes at the price of a computationally expensive pre-processing of the entire corpus. Another specific issue in the biomedical domain is 'shallow annotation' in biomedical corpora. As mentioned before in Section 3.2.2, term frequencies and co-occurrences in the biomedical domain are often gathered based on their usage in annotating gene mentions in a corpus. A potential issue as noted by Sevilla *et al.* (2005) is that annotators sometimes use more general concepts for annotation even if a more specific concept is more suitable, possibly by mistake or due to their lack of knowledge. As a result, the statistics can be biased toward more general concepts, leading to spurious prediction in methods that depends on corpus statistics (e.g., some IC-based methods and distributional similarity methods). Thus, the authors suggested that methods that are only based on the topological structure of a semantic graph (e.g., path-based methods) should be a better alternative in this domain.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

6.3 Purpose of the task

Another aspect to take into account is the purpose of the task. As discussed before, methods that are particularly tailored for measuring semantic similarity can be found to have inferior performance in assessing general relatedness. They may be better options for tasks such as synonym detection and taxonomy learning, but less effective for WSD or NED where relatedness are more important. Likewise, methods that assess relatedness may also produce spurious predictions of similarity and become unsuitable for some tasks. The lexical units (e.g., words or entities) involved in a task will also have an impact on the choice of underlying background information resources due to different focuses of knowledge in these resources.

For tasks built on top of lexical semantic relatedness, the trade-off between accuracy and complexity of the methods will be a major factor, particularly in large scale tasks, since computing semantic relatedness adds extra pre-processing cost. To our knowledge, no work has studied the complexity of semantic relatedness methods, and there is only limited work on comparing applications of lexical semantic relatedness. Curran and Moens (2002) compared several distributional similarity metrics in a thesaurus generation task. Budanitsky and Hirst (2006) compared five WordNet-based methods in a malapropism application, while Patwardhan and Pedersen (2006) performed similar studies in a WSD task. Bollegala *et al.* (2007) compared their method against Sahami and Heilman (2006) and Chen *et al.* (2006) in an entity clustering task. Zesch and Gurevych (2010a) and Tsatsaronis *et al.* (2010) carried out comparative evaluations of several methods in a word choice application. Where *in-vitro* evaluation results are also available in these studies, there is no strong evidence of a positive correlation between the accuracies of semantic relatedness methods (as in *in-vitro* evaluation) and their contribution to the application (as in *in-vivo* evaluation). For example, Patwardhan and Pedersen's (2006) method obtained an improvement of 0.15 point in correlation over the Jiang and Conrath's (1997) method on the RG65 dataset (Rubenstein and Goodenough 1965); however, it was outperformed by this method when applied to a WSD task. Similarly, Zesch and Gurevych (2010a) showed that several better performing methods in the *in-vitro* evaluation achieved lower accuracies in a word choice application.

Given these observations, one may want to reconsider their requirements for the accuracies of semantic relatedness methods when choosing one for their applications; particularly when dealing with a large amount of data such as in WSD or sense clustering.

7 Tools for lexical semantic relatedness

Due to the importance of lexical semantic relatedness and their frequent use in numerous applications, a number of tools have been developed and made available for research and application purposes.

WordNet::Similarity (Pedersen, Patwardhan and Michelizzi 2004) is a Perl software package to compute semantic similarity and relatedness between a pair of WordNet concepts. It implements six measures of similarity, and three measures of relatedness. The similarity measures include the shortest path-based measure similar to Rada

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

et al. (1989), Wu and Palmer (1994), Resnik (1995), Jiang and Conrath (1997), Leacock and Chodorow (1998), and Lin (1998b). The relatedness measures include Hirst and St-Onge (1998), Banerjee and Pedersen (2003), and Patwardhan and Pedersen (2006). The tool was later reimplemented in Java by Hope (2008) and Shima (2011).

Ponzetto and Strube (2007) released a Java API for computing semantic relatedness between words using the Wikipedia category tree. Given a pair of words, the tool retrieves the Wikipedia concepts (i.e., articles) they possibly refer to and computes relatedness between concepts using several methods previously reported in Strube and Ponzetto (2006), including Lesk (1986), Rada *et al.* (1989), Wu and Palmer (1994), Resnik (1995), Leacock and Chodorow (1998), Banerjee and Pedersen (2003), and Seco *et al.* (2004). The API also provides methods for accessing Wikipedia content.

In addition, several tools are developed for accessing Wikipedia content. Parse::MediaWikiDump (Riddle 2006) is a Perl module for parsing the Wikipedia XML dump to access the article contents. WikiPrep (Gabrilovich 2007) is a Perl script that processes the Wikipedia XML dump into different format that makes it easier to access the content and structural elements of Wikipedia. JWPL:Java-based WikiPedia Library (Zesch, Müller and Gurevych 2008a) is a Java API that converts Wikipedia or Wiktionary dump into relational databases and provides object-oriented representation of and access to Wikipedia and Wiktionary content.

Concerning dataset generation, Zesch and Gurevych (2006) introduced the DEXTRACT tool, a semi-automatic corpus-based approach for creating evaluation datasets for lexical semantic relatedness. The tool takes a corpus as input and automatically generates word pairs, which are then filtered and selected based on several constraints. The dataset is then published to users through a graphical user interface (GUI) interface, where human judgements are collected and gold standard can be created.

In the biomedical domain, the majority of tools are developed based on the GO. A summary of these, including the implemented methods and resources, can be found in Pesquita *et al.* (2009). Concerning other background resources, McInnes *et al.* (2009) developed UMLS-Similarity, an open-source Perl package that calculates semantic similarity between biomedical concepts using the UMLS Metathesaurus as the knowledge base. Li *et al.* (2011) developed DOSim, an R-based package for computing semantic similarity between disease terms based on disease ontology. Both have implemented several state-of-the-art measures of semantic similarity, such as Rada *et al.* (1989), Wu and Palmer (1994), Couto *et al.* (2005), and Al-Mubaid and Nguyen (2006).

8 Conclusion and future directions

This paper has reviewed existing methods for measuring lexical semantic relatedness. We have summarized different background information resources used by the literature, and discussed different categories of methods covering both the general and biomedical domains based on their rationale. We have summarized the evaluation

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

methodologies, and presented an analysis based on previously published resources. It has been shown that comparing semantic relatedness methods is a nontrivial task. Firstly, different methods have been evaluated on different datasets, using different versions of background information resources, and different evaluation metrics. All of these have made a comprehensive comparison rather difficult. Secondly, the results of the comparison are inconclusive – no single method can consistently outperform others on all datasets, although generally vector-based methods tend to have more balanced performances. The conclusion is that choosing semantic relatedness methods should depend on a number of inter-related factors. Each category of method has some advantages over others but equally suffers from certain limitations. These limitations are often bound to the underlying background information resources used by the methods. On the other hand, the choices of semantic relatedness methods are often limited by the availability of background information resources in relevant domains. Finally, from the application point of view, with limited data it is unclear whether improvement in the accuracies of semantic relatedness methods always leads to positive and proportional improvement in the application built on top of it. For this reason, it is often necessary to balance the trade-off between the potential accuracy of semantic relatedness methods and their complexity when choosing for applications. Furthermore, we identified remaining issues in this research field and suggest future research directions in the following.

Comparative evaluation. Despite the availability of benchmarking datasets for *in-vitro* evaluation, many studies did not report *in-vitro* experiments, or only used the most well-known datasets. This makes it difficult to compare their methods against others and possibly provides too partial a view of the capability of methods. Thus, it is important for future studies to carry out comparative evaluation to support meaningful comparisons with other work. Similarly, we encourage *in-vivo* experiments to use standard benchmarks, and new datasets should be published where possible to encourage experiment replication and comparison.

Domain generality and adaptability. Measuring lexical semantic relatedness is an important task in both general and specialized domains. We have discussed studies from both areas from a general point of view, and shown that they share common rationales. For this reason, adapting the methods across domains should be straightforward. Unfortunately, few studies have included a cross-domain evaluation. Therefore, we encourage future research to share lessons learnt from both communities and address domain adaptability in relevant evaluations.

Multilinguality. In theory, semantic relatedness methods can be applied to different languages by simply adapting the methods to background information resources of particular languages. In practice, although the performance of different methods for different languages are generally consistent, some may require specialization due to different characteristics (e.g., morphological, syntactical) of languages. For example, Gurevych and Niederlich (2005) showed that for strongly inflected languages such as German, methods that require counting word frequencies such as IC-based methods can be biased, and it is necessary to apply stemming or lemmatization to reduce inflected words to certain base forms to obtain more accurate predictions.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Such insights can be valuable references but they are currently rare due to lack of literature on multilingual semantic relatedness. Meanwhile, research on cross-lingual semantic relatedness can be particularly valuable for its broad application in cross-lingual Information Retrieval, cross-lingual document classification, and other related tasks.

Harnessing different background information resources. Different background information resources encode knowledge in different ways and have different focuses. On the one hand, we have discussed that this has biased some resources to particular tasks or categories of methods. On the other hand, Zhang *et al.* (2011) suggested that they may complement each other for evaluating lexical semantic relatedness. They have shown that by combining WordNet and Wikipedia under a uniform model, their method obtains competitive results and even achieved comparable results on biomedical datasets, for which state-of-the-art has to exploit domain-specific resources. Recently some methods (Gracia and Mena 2008; Agirre *et al.* 2009; Han and Zhao 2010; Szarvas *et al.* 2011) have explored a similar direction, and we believe that harnessing different background information resources can further improve the accuracy of semantic relatedness methods.

Opportunities with linked data. With increasing popularity of the Semantic Web, a new form of background information resource that can be potentially useful is Linked Data. Linked Data refers to the data, information, and knowledge interconnected and exposed on the Semantic Web (http://linkeddata.org/). A typical example of Linked Data is the DBPedia project (Bizer *et al.* 2009), a free online multimillion triple store initially created by extracting structured information (e.g., links, categories, infoboxes) from Wikipedia and representing knowledge as triples, which connects entities and concepts with relations. The true power of the Linked Data rests on the *linkedness of data*, bringing both opportunities and challenges to research in lexical semantic relatedness. On the one hand, information and knowledge covering unlimited domains and from heterogeneous sources are integrated and represented in a uniform way, granting easier access to the ever largest knowledge base. On the other hand, finding only the relevant information efficiently and effectively in this unlimited open *linkedness* is a challenging task (Gangemi and Presutti 2010). We believe that another interesting direction is investigating methods of using Linked Data as background information resource for evaluating lexical semantic relatedness. It will be particularly valuable to study how relevant knowledge can be effectively discovered, selected, and formalized to support the task.

Acknowledgments

Part of this research has been funded under the EC 7th Framework Program, in the context of the SmartProducts project (231204).

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'09), pp. 19–27. Stroudsburg, PA, USA: Association for Computational Linguistics.

Al-Mubaid, H., and Nguyen, H. 2006. A cluster-based approach for semantic similarity in the biomedical domain. In *Proceedings of the 28th International Conference of IEEE Engineering in Medicine and Biology Society*, New York, USA, August 30–September 3, pp. 2713–7.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17): 3389–402.

Alvarez, M., and Liam, S. 2007. A graph modeling of semantic similarity between words. In *Proceedings of the International Conference on Semantic Computing* (ICSC'07), pp. 355–62. Washington, DC, USA: IEEE Computer Society.

Banerjee, S., and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 805–10. San Francisco, CA, USA: Morgan Kaufmann.

Bär, D., Zesch, T., and Gurevych, I. 2011. A reflective view on text similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011* (RANLP 2011), Hissar, Bulgaria, pp. 515–20.

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

Bhattacharya, A., Bhowmick, A., and Singh, A. 2010. Finding top-k similar pairs of objects annotated with terms from an ontology. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management* (SSDBM'10), pp. 214–32. Berlin, Germany: Springer-Verlag.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. 2009. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics* **7**(3), 154–65.

Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. An integrated approach to measuring semantic similarity between words using information available on the web. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 340–7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. 2007. UniProtKB/Swiss-Prot. *Methods in Molecular Biology* **406**, 89–112.

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

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. 2004. The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Research* **32**(Database), D262–6.

Chen, H., Lin, M., and Wei, Y. 2006. Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1009–16. Stroudsburg, PA, USA: Association for Computational Linguistics.

Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. 1998. SGD: saccharomyces genome database. *Nucleic Acids Research* **26**(1), 73–9.

Chinchor, N. 2001. *Message Understanding Conference (MUC) 7*. LDC2001T02, Philadelphia, Penn: Linguistic Data Consortium.

Chinchor, N., and Sundheim, B. 2003. *Message Understanding Conference (MUC) 6*. LDC Catalog No.: LDC2003T13. Philadelphia, PA: Linguistic Data Consortium.

Cilibrasi, R., and Vitanyi, P. 2007. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* **19**(3), 370–83.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Collins, A., and Loftus, E. 1975. A spreading-activation theory of semantic processing. *Psychological Review* **82**(6), 407–28.

Couto, F., Silva, M., and Coutinho, P. 2005. Semantic similarity over the Gene Ontology: family correlation and selecting disjunctive ancestors. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (CIKM'05), pp. 343–4. New York, NY, USA: ACM.

Cramer, I., and Finthammer, M. 2008. An evaluation procedure for WordNet-based lexical chaining: methods and issues. In *Proceedings of the 4th Global WordNet Meeting*, pp. 120–46. Szeged, Hungary: University of Szeged.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 708–16. Stroudsburg, PA, USA: Association for Computational Linguistics.

Curran, J., and Moens, M. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition* (ULA'02), pp. 59–66. Stroudsburg, PA, USA: Association for Computational Linguistics.

Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcntara, R., Darsow, M., Guedj, M., and Ashburner, M. 2007. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research* **36**(Database), D344–50.

Dolan, B., Quirk, C., and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING'04), pp. 350–6. Stroudsburg, PA, USA: Association for Computational Linguistics.

Egozi, O., Markovitch, S., and Gabrilovich, E. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions of Information Systems* **29**(2), 8:1–8:34.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press.

Finkelstein, F., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. 2002. Placing search in context: the concept revisited. *ACM Transactions of Information Systems* **20**(1), 116–31.

Firth, J. R. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis* (special volume of the Philological Society), pp. 1–32. Harlow, UK: Longman.

Gabrilovich, E. 2007. Wikipedia preprocessor (WikiPrep). http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/#references. Accessed March 16, 2012).

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence* (IJCAI'07), pp. 1606–11. San Francisco, CA, USA: Morgan Kaufmann.

Gangemi, A., and Presutti, V. 2010. Towards a pattern science for the semantic web. *Emantic Web Journal* **1**(1–2), 61–8.

Gentleman, R. (2005). Visualizing and distances using GO. http://bioconductor.org/packages/2.0/bioc/vignettes/GOstats/inst/doc/GOvis.pdf. Accessed March 16, 2012.

Gouws, S., van Rooyen, G-J, and Engelbrecht, H. A. 2010. Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. In *Proceedings of the COLING 2010, 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, Beijing, China, pp. 46–54.

Gracia, J., and Mena, E. 2008. Web-based measure of semantic relatedness. In *Proceedings of the 9th International Conference on Web Information Systems Engineering (WISE'08)*, pp. 136–150. Berlin, Germany: Springer-Verlag.

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*, pp. 767–78. Berlin, Germany: Springer-Verlag.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Gurevych, I., and Niederlich, H. 2005. Computing semantic relatedness in German with revised information content metrics. In *Proceedings of ÖntoLex 2005 – Ontologies and Lexical Resources (ÏJCNLP'05) Workshop*, pp. 28–33. Berlin, Germany: Springer-Verlag.

Halavais, A., and Lackaff, D. 2008. An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication* **13**(2), 429–40.

Han, X., and Zhao, J. 2010. Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 50–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Harman, D., and Liberman, M. 1993. *TIPSTER* vol. 1. Philadelphia, PA, USA: Linguistic Data Consortium.

Harrington, B. 2010. A semantic network approach to measuring relatedness. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 356–64. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Haveliwala, T. 2002. Topic-sensitive PageRank. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*, pp. 517–26. New York, NY, USA: ACM.

Hirst, G., and St-Onge, D. 1998. Lexical chains as representation of context for the detection and correction malapropisms. In C. FellBaum (ed.), *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, pp. 305–32. Cambridge, MA, USA: MIT Press.

Holloway, T., Bozicevic, M., and Börner, K. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Journal of Complexity, Special issue on Understanding Complex Systems* **12**(3), 30–40.

Hope, D. 2008. *Java WordNet::Similarity (beta)*. http://www.sussex.ac.uk/Users/ drh21/. Accessed March 16, 2012.

Hughes, T., and Ramage, D. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 581–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hunter, S., Apweiler, R., Attwood, K., Bairoch, A., Bateman, A., Binns, D., Bork, P., and Das, U. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**(Database), D211–5.

Jarmasz, M., and Szpakowicz, S. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, Borovets, Bulgaria, September 10–12, pp. 212–9.

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

Jones, K. 1973. Index term weighting. *Information Storage and Retrieval* **9**(11), 619–33.

Kanehisa, M., and Goto, S. 2006. KEGG: Kyoto encyclopedia of genes and genomes. *Artificial Intelligence* **28**(1), 27–30.

Kilgarriff, A. 2007. Googleology is bad science. *Journal of Computational Linguistics* **33**(1), 147–51.

Kliegr, T., Chandramouli, K., Nemrava, J., Svatek, V., and Izquierdo, E. 2008. Combining image captions and visual analysis for image concept classification. In *Proceedings of the 9th International Workshop on Multimedia Data Mining Held in Conjunction with the ACM SIGKDD 2008 (MDM'08)*, pp. 8–17. New York, NY, USA: ACM.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Kohler, S., Schulz, M., Krawitz, P., Bauer, S., Dolken, S., Ott., C., Mundlos, C., Horn, C., Horn, D., Mundlos, S., and Robinson, P. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics* **85**(4), 457–64.

Kozima, H., and Furugori, T. 1993. Similarity between words computed by spreading activation on an English dictionary. In *Proceedings of the 6th Conference on European Chapter of the Association for Computational Linguistics (EACL '93)*, pp. 232–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kucera, H., and Francis, W. 1967. *Computational Analysis of Present-Day American English*. Providence, RI, USA: Brown University Press.

Kunze, C., and Lemnitzer, L. 2002. GermaNet – representation, visualization, application. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Spain, pp. 1485–91. Paris, France: ELRA.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In C. FellBaum (ed.), *WordNet: An Electronic Lexical Database*, pp. 305–32. Cambridge, MA, USA: MIT Press.

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (ACL'99)*, pp. 25–32. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lee, J., Kim, M., and Lee, Y. 1993. Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation* **49**(2), 188–207.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task (CONLL Shared Task '11)*, pp. 28–34. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lee, M., Pincombe, B., and Welsh, M. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 1254–9. Chicago, USA: Lawrence Erlbaum.

Lei, Z., and Dai, Y. 2006. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics* **7**, 491.

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)*, pp. 24–6. New York, NY, USA: ACM.

Li, Y., Bandar, Z., and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* **15**(4), 871–82.

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., and Li, X. 2011. DOSim: an R package for similarity between diseases based on disease ontology. *BMC Bioinformatics* **12**, 266.

Li, L., Hu, X., Hu, B., Wang, J., and Zhou, Y. 2009. Measuring sentence similarity from different aspects. In *Proceedings of the 8th International Conference on Machine Learning and Cybernetics (ICMLC 2009)*, Baoding, China, pp. 2244–9.

Li, Y., McLean, D., Bandar, Z., O'Shea, J., and Crockett, K. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering* **18**(8), 1138–50.

Li, B., Wang, J., Feltus, F., Zhou, J., and Luo, F. 2010. Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. In *Proceedings of the 11th International Conference on Bioinformatics and Computational Biology*, pp. 166–72. Las Vegas, NV, USA: CSREA Press.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Lin, D. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98)*, pp. 768–74. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, D. 1998b. An information-theoretic definition of similarity. In *Proceedings of the 5th International Conference on Machine Learning, (ICML '98)*, pp. 296–304. San Francisco, CA, USA: Morgan Kaufmann.

Liu, H., and Chen, Y. 2010. Computing semantic relatedness between named entities using Wikipedia. In *Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI '10)*, pp. 388–92. Washington, DC, USA: IEEE Computer Society.

Liu, X., Zhou, Y., and Zheng, R. 2007. Measuring semantic similarity in Wordnet. In *Proceedings of the 6th International Conference on Machine Learning and Cybernetics*, pp. 3431–5. New York, NY, USA: IEEE.

Lord, P., Stevens, R., Brass, A., and Goble, C. 2003a. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**(10), 1275–83.

Lord, P., Stevens, R., Brass, A., and Goble, C. 2003b. Semantic similarity measures as tools for exploring the Gene Ontology. In *Proceedings of Pacific Symposium on Biocomputing*, Lihue, HI, USA, January 3–7, pp. 601–12.

Maguitman, A., Menczer, F., Roinestad, H., and Vespignani, A. 2005. Algorithmic detection of semantic similarity. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*, pp. 107–116. New York, NY, USA: ACM.

Marcus, M., Marcinkiewicz, M., and Santorini, B. 1993. Building a large annotated corpus of English: the Penn treebank. *Journal of Computational Linguistics* **19**(2), 313–30.

Matsuo, Y., Sakaki, T., Uchiyama, K., and Ishizuka, M. 2006. Graph-based word clustering using a web search engine. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, pp. 542–50. Stroudsburg, PA, USA: Association for Computational Linguistics.

McInnes, B., Pedersen, T., and Pakhomov, S. 2009. UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In *Proceedings of AMIA Annual Symposium*, San Francisco, CA, USA, November 4–18, pp. 431–5.

McKusick, V. 1998. *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*, 12th ed. Baltimore, MD: The Johns Hopkins University Press.

McQuilton, P., St. Pierre, S., Thurmond, J., and the FlyBase Consortium. 2011. FlyBase 101 – the basics of navigating flyBase. *Nucleic Acids Research* **39**, 1–9.

Meyer, C., and Gurevych, I. 2010. How web communities analyze human language: word senses in Wiktionary. In *Proceedings of the 2nd Web Science Conference,* Raleigh, NC, April 26–27.

Mihalcea, R., Corley, C., and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*, pp. 775–80. Palo Alto, CA,USA: AAAI Press.

Mihalcea, R., and Moldovan, D. 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)*, pp. 152–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Milne, D., Medelyan, O., and Witten, I. 2006. Mining domain-specific thesauri from Wikipedia: a case study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, (WI'06)*, pp. 442–8. Washington, DC, USA: IEEE Computer Society.

Milne, D., and Witten, I. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*, pp. 25–30. Palo Alto, CA, USA: AAAI Press.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, D., Grishman, R., Meyers, A., Brunstain, A., Ferro, L., and Sundheim, B. 2003. *TIDES Extraction (ACE) 2003 Multilingual Training Data*. LDC Catalog Number: LDC2004T09, pp. 25–30. Philadelphia, PA: Linguistic Data Consortium.

Mohler, M., and Mihalcea, R. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, pp. 567–75. Stroudsburg, PA, USA: Association for Computational Linguistics.

Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Journal of Computational Linguistics*, **17**(1), 21–48.

Morris, J., and Hirst, G. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics (CLS '04)*, pp. 46–51. Stroudsburg, PA, USA: Association for Computational Linguistics.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., Magistry, P., and Chu-Ren, H. 2009. Wiktionary and NLP: improving synonymy networks. In *Proceedings of the 2009 Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web '09)*, pp. 19–27. Stroudsburg, PA, USA: Association for Computational Linguistics.

Navigli, R. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44)*, pp. 105–12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Navigli, R. 2009. Word sense disambiguation: a survey. *ACM Computing Survey* **41**(2), 10:1–10:69.

Othman, R., Deris, S., and Illias, R. 2007. A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. *Journal of Biomedical Informatics* **41**(1), 529–38.

Pakhomov, S., Coden, A., and Chute, C. 2004. Creating a test corpus of clinical notes manually tagged for part-of-speech information. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*, pp. 62–5. Geneva, Switzerland: Association for Computational Linguistics.

Pakhomov, S., Mcinnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. 2010. Semantic similarity and relatedness between clinical terms: an experimental study. *Proceedings of AMIA 2010 Symposium*, 572–6. Washington, DC, USA: American Medical Informatics.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A., and Vyas, V. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pp. 938–47. Berlin, Germany: Association for Computational Linguistics.

Patwardhan, S., and Pedersen, T. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Proceedings of the EACL 2006 Workshop Making Sense of Sense – Bringing Computational Linguistics and Psycholinguistics Together*, pp. 1–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* **40**(3), 288–99.

Pedersen, T., Patwardhan, S., and Michelizzi, J. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004 (HLT-NAACL–Demonstrations '04)*, pp. 38–41. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pekar, V., and Staab, S. 2002. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th International Conference on Computational Linguistics – vol. 1, (COLING'02)*, pp. 1–7. Stroudsburg, PA, USA: Association for Computational Linguistics.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Pesquita, C., Faria, D., Falcão, A., Lord, P., and Couto, F. 2009. Semantic similarity in biomedical ontologies. *PLoS Computational Biology* **5**(7):e1000443. 1–12.

Petrakis, E., Varelas, G., Hliaoutakis, A., and Raftopoulou, P. 2006. Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. In *Proceedings of the 4th Workshop on Multimedia Semantics (WMS'06)*, Chania, Crete, June 19–21, pp. 44–52.

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

Pirró, G., and Seco, N. 2008. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II: On the Move to Meaningful Internet Systems (OTM '08)*, pp. 1271–88. Berlin, Germany: Springer-Verlag.

Ponzetto, S., and Strube, M. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL '06)*, pp. 192–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ponzetto, S., and Strube, M. 2007. An API for measuring the relatedness of words in Wikipedia. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*, pp. 49–52. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ponzetto, S., and Strube, M. 2011. Taxonomy induction based on a collaboratively built knowledge repository. *Journal of Artificial Intelligence* **175**(9–10), 1737–1756.

Pozo, A., Pazos, F., and Valencia, A. 2008. Defining functional distances over Gene Ontology. *BMC Bioinformatics* **9**, 50.

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

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pp. 337–46. New York, NY, USA: 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 (IJCAI'95)*, pp. 448–53. San Francisco, CA, USA: Morgan Kaufmann.

Richardson, R., and Smeaton, A. 1995. Using WordNet in a knowledge-based approach to information retrieval. Technical Report CA-0196, School of Computer Applications, Dublin City University.

Riddle, T. 2006. Parse::MediaWikiDump. http://search.cpan.org/~triddle/Parse-MediaWikiDump-1.0.6/lib/Parse/MediaWikiDump.pm. Accessed March 16, 2012.

Riensche, R., Baddeley, B., Sanfilippo, A., Posse, C., and Gopalan, B. 2007. XOA: web-enabled cross-ontological analytics. In *Proceedings of the 1st International Workshop on Service-Oriented Technologies for Biological Databases and Toolsat in the ICWS/SCC Conference*, pp. 99–105. Washington, DC, USA: IEEE Computer Society.

Rodríguez, M., and Egenhofer, M. 2003. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering* **15**(2), 442–56.

Rose, T., Stevenson, M., and Whitehead, M. 2002. The Reuters corpus volume 1-from yesterdays news to tomorrows language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 29–31. Paris, France: ELRA.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

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

Ruiz-Casado, M., Alfonseca, E., and Castells, P. 2005. Using context-window overlapping in synonym discovery and ontology extension. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Sahami, M., and Heilman, T. 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)*, pp. 377–86. New York, NY, USA: ACM.

Schickel-Zuber, V., and Faltings, B. 2007. OSS: a semantic similarity function based on hierarchical ontologies. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*, pp. 551–6. San Francisco, CA, USA: Morgan Kaufmann.

Schlicker, A., Domingues, F., Rahnenführer, J., and Lengauer, T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**, 302.

Seco, N., Veale, T., and Hayes, J. 2004. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, Valencia, Spain, August 22–27, pp. 1089–90.

Sevilla, J., Segura, V., Podhorski, A., Guruceaga, E., Mato, J., Martinez-Cruz, L., Corrales, F., and Rubio, A. 2005. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(4), 330–8.

Sheng, H., Chen, H., Yu, T., and Feng, Y. 2010. Linked data-based semantic similarity and data mining. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2010)*, pp. 104–8. New York, NY: IEEE Systems, Man, and Cybernetics Society.

Shima, H. 2011. WS4J. http://code.google.com/p/ws4j/. Accessed March 16, 2012.

Speer, N., Spieth, C., and Zell, A. 2004. A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, October 7–8, pp. 252–9. New York, NY, USA: IEEE.

Staab, S., Braun, C., Bruder, I., Düsterhöft, A., Heuer, A., Klettke, M., Neumann, G., Prager, B., Pretzel, J., Schnurr, H., Studer, R., Uszkoreit, H., and Wrenger, B. 1999. GETESS: searching the web exploiting German texts. In *Proceedings of the 3rd International Conference on Cooperative Information Agents III (CIA'99)*, pp. 113–24. Berlin, Germany: Springer-Verlag.

Strube, M., and Ponzetto, S. 2006. WikiRelate! computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)*, pp. 1419–24. Palo Alto, CA, USA: AAAI Press.

Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM '93)*, pp. 67–74. New York, NY, USA: ACM.

Szarvas, G., Zesch, T., and Gurevych, I. 2011. Combining heterogeneous knowledge resources for improved distributional semantic models. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, pp. 289–303. Tokyo, Japan: Springer-Verlag.

The BNC Consortium. 2007. The British National Corpus, Version 3 (BNC XML edition). http://www.natcorp.ox.ac.uk/. Accessed March 16, 2012. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

The Gene Ontology Consortium. 2005. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25–9.

Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. 2010. Text relatedness based on a word thesaurus. *Journal of Artificial Intelligence Research* **37**(1), 1–40.

Turdakov, D., and Velikhov, P. 2008. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. *Proceedings of*

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

the Spring Young Researcher's Colloquium On Database and Information Systems (CEUR workshop proceedings), St. Petersburg, Russia. Available at CEUR-WS.org.

Turney, P., and Pantel, P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* **37**, 141–88.

Tversky, A. 1977. Features of similarity. *Psychological Review* **84**(4), 327–52.

Vapnik, V. 1998. *Statistical Learning Theory*. Chichester, UK: Wiley.

Wang, J., Du, Z., Payattakool, R., Yu, P., and Chen, C. 2007. A new method to measure the semantic similarity of GO terms. *BMC Bioinformatics* **23**(10), 1274–81.

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

Weeds, E. 2003. *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, University of Sussex.

Wojtinnek, P., and Pulman, S. 2011. Semantic relatedness from automatically generated semantic networks. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS '11)*, pp. 390–4. Oxford, UK: Association for Computational Linguistics.

Wu, Z., and Palmer, M. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL '94)*, pp. 133–8. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. 2005. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Research* **33**(9), 2822–37.

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

Yang, D., and Powers, D. 2005. Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the 28th Australasian Conference on Computer Science (ACSC '05)*, pp. 315–22. Darlinghurst, Australia: Australian Computer Society.

Yang, D., and Powers, D. 2006. Verb similarity on the taxonomy of Wordnet. In *Proceedings of the 3rd International WordNet Conference (GWC-06)*. Masaryk, Czech Republic: Masaryk University.

Yang, X., and Su, J. 2007. Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 528–35. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yazdani, M., and Popescu-Belis, A. 2010. A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks. In *Proceedings of the 2010 IEEE 4th International Conference on Semantic Computing (ICSC '10)*, pp. 424–9. Washington, DC, USA: IEEE Computer Society.

Ye, P., Peyser, B., Pan, X., Boeke, J., Spencer, F., and Bader, J. 2005. Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular Systems Biology* 1:2005.0026. pp. 1–12.

Yeh, E., Ramage, D., Manning, C., Agirre, E., and Soroa, A. 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the ACL 2009 Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-4)*, pp. 41–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yu, H., Gao, L., Tu, K., and Guo, Z. 2005. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene* **352**, 75–81.

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

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press

Zesch, T., and Gurevych, I. 2007. Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, pp. 1–8 Stroudsburg, PA, USA: Association for Computational Linguistics.

Zesch, T., and Gurevych, I. 2010a. Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words. *Natural Language Engineering* **16**(1), 25–59.

Zesch, T., and Gurevych, I. 2010b. The more the better? Assessing the influence of Wikipedia's growth on semantic relatedness measures. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*. Paris, France: European Language Resources Association (ELRA).

Zesch, T., Müller, C., and Gurevych, I. 2008a. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pp. 1646–52. Paris, France: European Language Resources Association (ELRA).

Zesch, T., Müller, C., and Gurevych, I. 2008b. Using Wiktionary for computing semantic relatedness. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI'08)*, pp. 861–6. Palo Alto, CA, USA: AAAI Press.

Zhang, Z., Gentile, A., and Ciravegna, F. 2011. Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pp. 991–1002. Stroudsburg, PA, USA: Association for Computational Linguistics.

Ziegler, C., Simon, K., and Lausen, G. 2006. Automatic computation of semantic proximity using taxonomic knowledge. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*, pp. 465–74. New York, NY, USA: ACM.

https://doi.org/10.1017/S1351324912000125 Published online by Cambridge University Press