A survey of semantic relatedness evaluation datasets and procedures

Article *in* Artificial Intelligence Review · December 2019

DOI: 10.1007/s10462-019-09796-3

Mohamed Ali Hadj Taieb
Faculty of Sciences University of Sfax Tunisia
42 PUBLICATIONS **328** CITATIONS

Torsten Zesch
University of Duisburg-Essen
102 PUBLICATIONS **1,911** CITATIONS

Mohamed Ben Aouicha
University of Sfax
51 PUBLICATIONS **356** CITATIONS

Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09796-3

A survey of semantic relatedness evaluation datasets and procedures

Mohamed Ali Hadj Taieb[1] · Torsten Zesch[2] · Mohamed Ben Aouicha[1]

Abstract
Semantic relatedness between words is a core concept in natural language processing. While countless approaches have been proposed, measuring which one works best is still a challenging task. Thus, in this article, we give a comprehensive overview of the evaluation protocols and datasets for semantic relatedness covering both intrinsic and extrinsic approaches. One the intrinsic side, we give an overview of evaluation datasets covering more than 100 datasets in 20 different languages from a wide range of domains. To provide researchers with better guidance for selecting suitable dataset or even building new and better ones, we describe also the construction and annotation process of the datasets. We also shortly describe the evaluation metrics most frequently used for intrinsic evaluation. As for the extrinsic side, several applications involving semantic relatedness measures are detailed through recent research works and by explaining the benefit brought by the measures.

Keywords Semantic relatedness · Semantic similarity · Evaluation dataset · Evaluation metric · Evaluation procedure

1 Introduction

The measurement of the Semantic Relatedness (SR) between concepts or words is an important fundamental research topic in natural language processing. The basic idea is to quantify how close or distant two words are, e.g. *car* and *street* are usually perceived as more related than *car* and *banana*.

Measuring semantic relatedness has applications in many fields including: information retrieval (Akmal et al. 2014; Chen et al. 2017; Gurevych et al. 2007; Lopez-Gazpio et al.

✉ Mohamed Ali Hadj Taieb
mohamedali.hadjtaieb@gmail.com

Torsten Zesch
torsten.zesch@uni-due.de

Mohamed Ben Aouicha
mohamed.benaouicha@fss.usf.tn

[1] Faculty of Sciences of Sfax, Sfax University, Sfax, Tunisia

[2] Language Technology Lab, University of Duisburg-Essen, Duisburg, Germany

Published online: 23 December 2019

2017; Srihari et al. 2000; Uddin et al. 2013), machine translation (Liu et al. 2007), ontology learning and alignment (Sánchez and Moreno 2008; Jiang et al. 2014), opinion aspect extraction (Lin et al. 2016), plagiarism detection (Franco-Salvador et al. 2016), spelling correction (Budanitsky and Hirst 2006), and word sense disambiguation (Ben Aouicha et al. 2016a; Patwardhan et al. 2003). Through the continuous efforts of researchers, the methods for the analysis of semantic relatedness measures have been constantly improved, and many new approaches are emerging, especially related to word embeddings (Mikolov et al. 2013a; Pennington et al. 2014; Lastra-Díaz et al. 2019b).

Lastra-Díaz et al. (2019b) conducted a large reproducible survey on word embeddings and ontology-based methods for word similarity. They provide publicly the data (Lastra-Díaz et al. 2019c, a) generated through the experiments using different computing models.

Thus, it is very important to carefully examine the evaluation protocol consisting of the datasets and metrics in order to provide researchers with guidelines for selecting the best methods or developing even better ones. To give a quick overview that will be detailed later in this article, methods for quantifying the semantic relatedness between words are usually evaluated by comparing the computational results with human judgments. For example, humans might perceive *car/street* as having a high semantic relatedness (e.g. 0.9) and the computational method is supposed to closely approximate this result.

In this article, we give a comprehensive overview of the methodology of evaluating semantic relatedness measures. We focus on single word concepts and do not explicitly cover the at least equally large body of work on phrase or text similarity (Cer et al. 2017). We also do not cover approaches for computing semantic relatedness, as we focus exclusively on the evaluation side.[1] To our knowledge, the current study is the first that focuses on the assessment protocol and includes a detailed survey covering the datasets in several languages and covering cross-lingual relatedness. Moreover, these datasets are detailed according to several criteria mainly divided on two steps: word selection and annotation process. Our study also includes a list of metrics for measuring the performance of semantic relatedness approaches. An up-to-date list of the more than 100 datasets from 20 languages is maintained by the authors and publicly available.[2]

2 Terminology

Before we can further look at the different types of evaluation protocols used in the field, we have to talk about terminology. The terms *relatedness*, *distance*, *similarity*, or *association* (with or without the *semantic* prefix) are sometimes used interchangeably causing unnecessary confusion.

2.1 Semantic distance, similarity and relatedness

We use the term *semantic relatedness* in the general sense of *semantic proximity* or *semantic association*, i.e. how much connection humans perceive between two concepts. We make no assumptions regarding the causes of the perceived connection. This means *relatedness* includes *similarity*, a more specific case, where the sense of relatedness is dependent on the ‘degree of synonymy’, i.e. the amount of shared properties. For example, *car/*

[1] Surveys on semantic relatedness approaches are e.g. Feng et al. (2017), Harispe et al. (2015), Zhang et al. (2012).

[2] https://github.com/MohamedAliHadjTaieb/Semantic-measure-assessment-review-study.

plane are both means of transportation, both machines, both use fuel, have engines and wheels, etc. Therefore, the two concepts are considered semantically similar. Concepts that are semantically related are not necessarily similar, such as *car/street*.

Semantic distance is the inverse of *relatedness*, i.e. two very related concepts are very close (not distant), while two distant concepts, are not very related. We consider relatedness to be the more natural definition and will use it exclusively in this article.

Another common mistake in thinking about those concepts is to consider to early which kind of methods will be used for quantifying the relatedness afterwards. Distributional methods (Weeds 2003) are often connected to semantic relatedness, while knowledge bases and ontologies such as WordNet (Fellbaum 1998) are connected to Semantic Similarity (SS). However, we argue that the evaluation process should be as theory-neutral as possible in the first place. For example, distributional methods have shown remarkable results on semantic similarity tasks (Hadj Taieb et al. 2013). Thus, it is important to build the evaluation datasets in an appropriate way (discussed in Sect. 4) or to design applications that that make use of semantic distance measures (discussed in Sect. 6).

Ferdinand de Saussure states that the intralinguistic relations exist between words (de Saussure 1983). They are basically of two types: syntagmatic and paradigmatic. The syntagmatic relations are the relationships that a linguistic unit has with the units in the stretch of speech (the context) in which it occurs. The paradigmatic relations are the relations that a linguistic unit has with units by which it may be replaced : sets of synonyms, pairs of antonyms, lexico-semantic groups, etc. (to get has the synonymic set: to obtain, to receive, to gain, to acquire, etc.).

The distinction between syntagmatic and paradigmatic relations is conventionally indicated by horizontal and vertical presentation. As it is illustrated through Fig. 1, the paradigmatic relations influence the word semantic similarity due to the fact that the words can be substituted through the interchanging the set of synonyms. The syntagmatic relations express the context which is exploited for studying the distributional semantic allowing the estimation of semantic relatedness.

Sahlgren (2006) states that "words have a syntagmatic relation if they co-occur, and a paradigmatic relation if they share same neighbors".

2.2 Multilingual and cross-lingual relatedness

The English community has so developed word similarity datasets, but, semantic representation for other languages has generally proved difficult to evaluate. Moreover, many works are conducted to provide datasets for evaluating the semantic similarity or relatedness measures in several languages other than the English (see Table 3). This is noticed as multinlingual semantic simialrity (Camacho-Collados et al. 2017). Some of these multilingual evaluation datasets are constructed on the basis of conventional English datasets such as Barzegar et al. (2018) and others are designed especially for the target language. Similarly to the case of multilingual datasets, the cross-lingual datasets have been constructed on the basis of conventional English word similarity datasets. As for the cross-lingual simialrity datasets are formed through word pairs pertaining to different languages as it is illustrated in Table 1.

The need for effective multilingual and cross-lingual text processing techniques is becoming increasingly important (Barzegar et al. 2018). Recently, multilingual embeddings that represent lexical items from multiple languages in a unified semantic space generating research attention and at the same time cross-lingual applications are studied (Franco-Salvador et al. 2016).

For instance, given the word *car* in English and the word *ruota* in Italian (En. *wheel*), a cross-lingual measure can return the relatedness of these two words despite the fact that they belong to two different languages.

Words are morphologically structured if they can be decomposed into multiple meaningful units (morphemes), as is the case with many words in English or more for other languages such as Arabic and Hindi. The morphologic problem concerns (1) the dataset construction process including in some cases automatic words selection step based on the analysis of dependant language corpus (2) and the pretreatment step before applying the semantic similarity/relatedness measure on the words pair of the datasets. The morphologic richness of some languages leads the researchers to specific treatments of the words pairs forming the datasets before transferring them as inputs to the semantic similarity/relatedness measures. These treatments include the stemmatisation for the distributional-based measures, the determination of the part of speech, transforming the word to the singular form, obtaining the canonical form , etc. For instance, segmentation of the affix *-er* from *whiter* or *farmer* leaves stems (*white* and *farm*) that are often, but not always, semantically similar to the original word. Other words, such as *corner* or *mother*, only appear to be morphologically structured; that is, they have a letter sequence, such as *er*, that functions as an affix in many other words, but not in these particular words. At the same time, their meaning is dissimilar to that of the words that are embedded in them (*corn* or *moth*). Ercan and Yildiz (2018), in their dataset AnlamVer, balance dataset word-pairs by their frequencies to evaluate the robustness of semantic models concerning out-of-vocabulary and rare words problems, which are caused by the rich derivational and inflectional morphology of the Turkish language. They talked about the words affecting the similarity measure such as the (i.e., unseen, out-of-vocabulary) or low occurrence (i.e., rare words) of a testing word in the training corpus. Distributional semantics community has been developing compositional models to overcome out-of-vocabulary and rare words problems. RW dataset (Luong et al. 2013) provides word frequency (rareness) based evaluation strategy to compositional model developers. Similarly, they aim to balance our dataset's word-pool by words' frequencies to assess generalization powers of such models. Konopik et al. (2017) talked about the manner for treating the morphology problem of the Czech language which is rich and highly irregular.

Measures of cross-language relatedness are useful for a large number of applications, including cross-language information retrieval (Nie et al. 1999; Monz and Dorr 2005), cross-language text classification (Gliozzo and Strapparava 2006), lexical choice in machine translation (Och and Ney 2000; Bangalore et al. 2007), induction of translation lexicons, cross-language annotation, resource projections to a second language (Riloff et al. 2002) and cross-language plagiarism detection (Franco-Salvador et al. 2016). Cross-lingual semantic similarity has been studied by several researches for different natural language processing applications such as word-to-word translation (Vulic and Moens 2014), comparing articles from different languages (Saad et al. 2014) or finding the semantic similarity of words from different languages (Vulic and Moens 2013), cross-lingual semantic textual similarity (Bjerva and Östling 2017), and cross-lingual link discovery (Narducci et al. 2017).

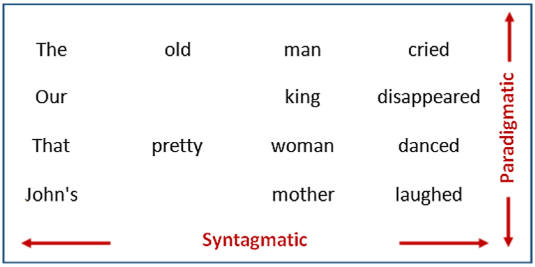


Fig. 1 Illustration of the syntagmatic and pragmatic relations

Table 1 Example pairs and their ratings (Camacho-Collados et al. 2017)

Monolingual			
DE	Tuberkulose	LED	0.25
ES	Zumo	Batido	3.00
EN	Multiple sclerosis	MS	4
IT	Nazioni Unite	Ban Ki-moon	2.25
FA	لئوناردوداوینچی	آخر شام	2.08
Cross-lingual			
DE–ES	Sessel	Taburete	3.08
DE–FA	Lawine	برف	2.25
DE–IT	Taifun	Ciclone	3.46
EN–DE	Pancreatic cancer	Chemotherapie	1.75
EN–ES	Jupiter	Mercurio	3.25
EN–FA	Film	گرایی پوچ	0.25
EN–IT	Island	Pensiola	3.08
ES–FA	Duna	بیابان	2.25
ES–IT	Estrella	Pianeta	2.83
IT–FA	Avvocato	نمایشگر	0.08

EN English, *DE* German, *ES* Spanish, *IT* Italian, *FA* Farsi

3 Intrinsic evaluation: datasets

A prerequisite for the intrinsic evaluation of semantic relatedness measures are appropriate datasets. They consist of a set of word pairs together with human judgments of their semantic similarity or relatedness value. Table 2 shows as an example the early dataset by Miller and Charles (1991). Datasets vary with respect to source of word pairs, type of annotated relationship, domain, covered part of speech, and language.

3.1 Dataset construction

As the size of an evaluation dataset is severely limited by the high costs of the manual annotation, words to be used in a dataset have to be carefully selected. Unfortunately, for most evaluation datasets information only partial information about the construction process is available. For our discussion, we thus highlight specific datasets that represent generic cases. Figure 2 illustrates the steps composing the process of the dataset construction based on the study elaborated in the present work.

3.1.1 Source of words

The exploited sources for extracting word pairs depend mainly on the domain and the language.

In SimLex999 (Hill et al. 2015), they focused on similarity as opposed to relatedness. In fact, they started with the 72,000 pairs of concepts in the University of South Florida Free Association Database (USF) (Nelson et al. 2004) dataset. They excluded pairs containing a multiple-word item *hot dog/mustard*, and those containing capital letter *Mexico/sun*. Then, they complement this dataset with entirely unassociated pairs. They paired up the concepts from the 900 associated pairs at random. From these random pairs, they excluded those that coincidentally occurred elsewhere in USF (having a higher degree of association).

The words constituting MEN3000 (Bruni et al. 2014) were randomly selected from words occurring at least 700 times in ukWaC and Wackypedia text corpora and at least 50 times as tags in the ESP-Game and MIRFLICKR-1M tagged image collections. In order to avoid selecting only pairs that were weakly related, they ranked all possible pairs by their cosines according to their text-based model Window.

MTurk287 (Radinsky et al. 2011) is created by applying a procedure by intersecting a set of all words in the New York Times news articles and a collection with entities extracted from DBpedia. They further proceed with removing rare words (words appearing less than 1000 over the entire time period). Next, for each word pair, the Point-wise Mutual Information (PMI) is computed using the whole articles. So, the obtained dataset includes both frequently and infrequently co-occurring words pertaining to the entire spectrum of co-occurrence values (as measured by mutual information).

Translating datasets The original English RG65 (Rubenstein and Goodenough 1965) and WordSim-353 (Finkelstein et al. 2002) datasets have been translated into other languages, either by experts (Gurevych 2005; Camacho-Collados et al. 2015; Joubarne and Inkpen 2011; Granada et al. 2014), or by means of crowdsourcing (Leviant and Reichart 2015), thereby creating equivalent datasets in languages other than English. Hassan et al. (2009) asked native speakers of Spanish, Romanian, and Arabic, who were also highly proficient in English, to translate the words in the two datasets. They assigned new annotations because the translation might change the meaning and the relatedness.

Barzegar et al. (2018) described SemR-11[3] a multi-lingual dataset for evaluating semantic similarity and relatedness for 11 languages. SemR-11 builds upon the English datasets MC30, RG65, WS353, and Simlex-999 and provide translated versions for all 11 languages. The word pairs were translated by paid professional translators, skilled in data localisation tasks.All translated pairs followed the protocol (1) Given a pair of words, translators should assume the most similar senses associated with the pair (2) Translators should preserve the lexical category of the sense identified for that word.

For the cross-lingual datasets, Kennedy and Hirst (2012) proposed a method which exploits two aligned monolingual word similarity datasets for the construction of a French-English cross-lingual dataset. Camacho-Collados et al. (2015) followed their initial idea and proposed a generalization of the approach which is exploited for automatically constructing cross-lingual datasets for any pair of languages. They created fifteen cross-lingual

[3] https://github.com/Lambda-3/Gold-Standards/tree/master/SemR-11.

datasets based on the RG65 datasets for English, French, German, Spanish, Portuguese, and Farsi. In SemEval-2017 Task2 (Camacho-Collados et al. 2017) Subtask 1, i.e. multilingual semantic similarity, has five datasets for the five languages English, Farsi, German, Italian, and Spanish. Based on these five datasets, 10 cross-lingual datasets were automatically generated for subtask 2, i.e. cross-lingual semantic similarity.

3.1.2 Pair construction

After suitable words have been selected, word pairs are formed. Randomly sampling word pairs yields unbalanced datasets with a lot of unrelated pairs and very few highly related pairs. Thus, several approaches have been proposed for controlling the distribution of relatedness in the resulting dataset.

Word co-occurrence For example pairs in the MTurk287 dataset (Radinsky et al. 2011) were sampled from frequently co-occurring words. Solely relying on this strategy is likely to result in a bias in favor of distributional methods that compute relatedness based on similar sources of knowledge.

Semantic relations Several filters have been proposed to obtain a more uniform distribution from related to unrelated pairs including enforcing a known semantic relation (Hypernym/Hyponym) between the words (Luong et al. 2013), separating similarity from association (Hill et al. 2015) to complement this dataset with entirely unassociated pairs using the University of South Florida Free Association Database (USF) (Nelson et al. 2004).

Combination Often the pair construction process involves a combination of the aforementioned factors. For example, the SimVerb3500 dataset (Gerz et al. 2016) is based on the USF norms dataset (Nelson et al. 2004) and VerbNet (Schuler 2005; Kipper et al. 2007). They extracted all possible verb pairs from USF using the associated POS tags available as part of USF annotations. Moreover, they exclude all USF pairs that had been associated by 2 or less participants in USF. Then they manually cleaned and simplified the list of pairs by removing all pairs according to already fixed properties such as the multi-word verbs (e.g. *give up*), the non-infinitive form of a verb (e.g. *accomplished* or *hidden*) and pairs containing at least one auxiliary verb (e.g. *be*).

3.1.3 Domain

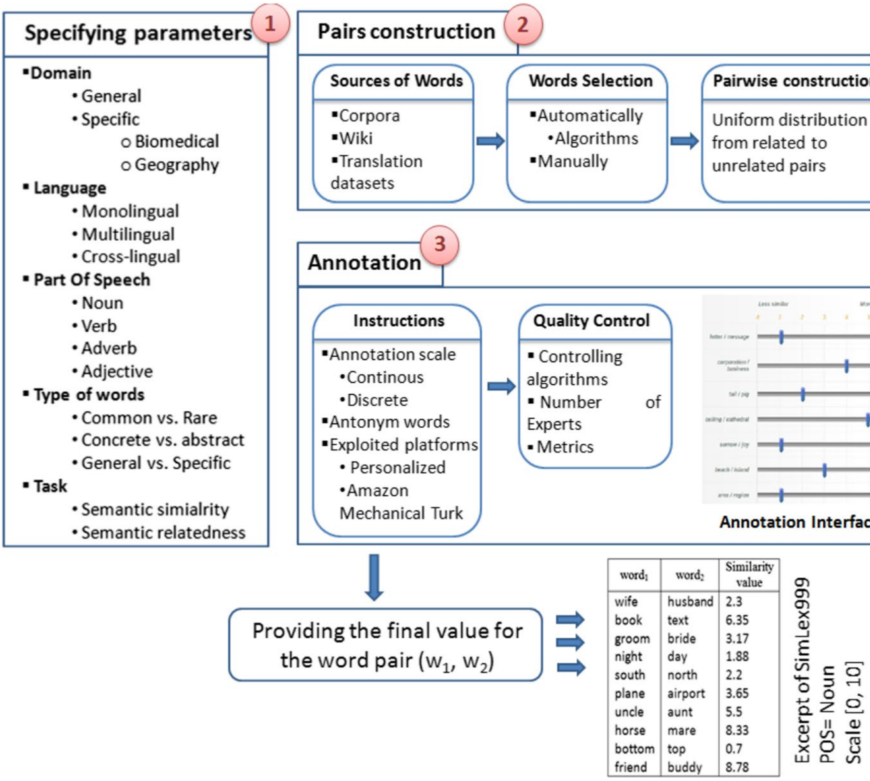
The majority of datasets are from the general domain, but there exist some datasets targeting specific domains such as biomedicine or geography. For example, Pakhomov et al. (2010) present a large dataset (UMNSRS724) for measuring semantic similarity and relatedness between biomedical concepts. Another example is the Geo Relatedness and Similarity Dataset (GeReSiD) (Ballatore et al. 2014) that contains geographic terms from the Open-Street Map project including both natural and man-made terms.

3.1.4 Type of words

Some datasets also control which type of words are used, e.g. common vs. rare (Luong et al. 2013), concrete vs. abstract, or general vs. specific. Another choice is whether single

Table 2 Example of a dataset (Miller and Charles 1991)

Word pair		Value	Word pair		Value
Car	Automobile	3.92	Lad	Brother	1.66
Gem	Jewel	3.84	Journey	Car	1.16
Journey	Voyage	3.84	Monk	Oracle	1.10
Boy	Lad	3.76	Cemetery	Woodland	0.95
Coast	Shore	3.70	Food	Rooster	0.89
Asylum	Madhouse	3.61	Coast	Hill	0.87
Magician	Wizard	3.50	Forest	Graveyard	0.84
Midday	Noon	3.42	Shore	Woodland	0.63
Furnace	Stove	3.11	Monk	Slave	0.55
Food	Fruit	3.08	Coast	Forest	0.42
Bird	Cock	3.05	Lad	Wizard	0.42
Bird	Crane	2.97	Cord	Smile	0.13
Tool	Implement	2.95	Glass	Magician	0.11
Brother	Monk	2.82	Rooster	Voyage	0.08
Crane	Implement	1.68	Noon	String	0.08



word$_1$	word$_2$	Similarity value
wife	husband	2.3
book	text	6.35
groom	bride	3.17
night	day	1.88
south	north	2.2
plane	airport	3.65
uncle	aunt	5.5
horse	mare	8.33
bottom	top	0.7
friend	buddy	8.78

Fig. 2 Generalisation of the dataset construction process. The annotation interface and the excerpt concern the dataset SimLex 999 (Hill et al. 2015)

words or multi-words (Li et al. 2013) are covered. There are also special datasets for named entities (Zie55) (Ziegler et al. 2006).

The majority of pairs in some datasets such as MEN3000 (Bruni et al. 2014) and RG65 (Rubenstein and Goodenough 1965) contain concrete items, although, the vast majority of adjective, noun and verb concepts in everyday language are abstract (Kiela et al. 2014). SimLex999 authors (Hill et al. 2015) aimed to include both concept types to facilitate the evaluation of models for both concrete and abstract concept meaning, and due to the cognitive and computational modeling differences between abstract and concrete concepts. They benefit from sampling pairs for SimLex999 from the USF dataset where most items have been rated according to concreteness on a scale of 1–7 by at least 10 human subjects. There is also clear variation in concreteness within each POS category. They, therefore, aimed to select pairs for SimLex-999 that covers different abstract-concrete levels for each POS category.

3.1.5 Part of speech

Most datasets only contain the same POS (N/N, V/V, A/A) in a single pair. There are only few cross POS datasets, although it is especially interesting for semantic relatedness e.g. in pairs like (*night*/*dark*) that are highly related (though not similar).

Another issue (especially for English) is POS ambiguity, e.g. *access* which can either be a noun or a verb. As this could lead to inconsistent ratings, a solution is to exclude pairs that don’t have a clear tendency towards a particular POS.

3.2 Dataset annotation

This process involves several details including the annotation methodology (crowd-sourcing vs. laboratory experiments), the annotation manual, the presentation of the word pairs to the annotators, the consistency assessment of the annotators to exclude non serious annotations, the target task (similarity or relatedness and monolingual or cross-lingual), the annotation scale (continuous or scalar scale) and whether the word context is taken into account.

3.2.1 Type of study

There are two main annotation methods: crowd-sourcing (Radinsky et al. 2011; Halawi et al. 2012; Panchenko et al. 2017; Camacho-Collados et al. 2017; Sakaizawa and Komachi 2017) and laboratory study (Miller and Charles 1991; Rubenstein and Goodenough 1965; Torsten and Iryna 2006).

Wang et al. (2015) conducted semantic transparency[4] rating experiments using both the traditional laboratory-based method and the crowd-sourcing-based method. Then they compared the rating data obtained from these two experiments. They observed very strong correlation coefficients for both overall semantic transparency rating data and constituent semantic transparency data (rho> 0.9) which means the two experiments may yield comparable data and crowd-sourcing-based experiment is a feasible alternative to the laboratory based experiment in linguistic studies.

[4] Semantic transparency is the degree to which the meaning of a compound word or an idiom can be inferred from its parts (or morphemes) (Bell and Schäfer 2016).The word *blueberry* is semantically transparent; the word *strawberry* is not.

The Russe dataset MJ contains 12,886 word pairs (Panchenko et al. 2017). These pairs have continuous relatedness scores. To estimate these scores they averaged 105 submissions of the shared task on Russian semantic similarity.

3.2.2 Annotation scale

In most of the existing datasets, the annotators were asked to assign a numeric score to each pair (e.g. 0–7 in SimLex999), and an average score ranking was computed. Note that a ranking of the pairs can be implicitly derived based on these average scores. This choice is probably due to the fact that a ranking of hundreds of pairs is an exhausting task for humans.

Avraham and Goldberg (2016) present a study for improving reliability of word similarity evaluation. They discussed mainly four problems related to the annotation process of the existing datasets which are (1) the problem is to ask the annotators for a numeric score while they advice to ask annotator for a ranking (2) the rating of different relations on the same scale leads to arbitrary decision of the annotators which affect the model score. So, they propose to give the instructions for assigning low scores to unpreferred-relation[5] pairs. (3) Rating different target words on the same scale is quite unnatural due to the fact that comparing between pairs that have different target-words, in contrast to pairs which share the target word, like (cat, pet) vs. (cat, animal). (4) Evaluation measure does not consider annotation decisions reliability which can be determined by the agreement of the annotators on this decision. Bruni et al. (2014) addressed the problem of rating scale by asking the annotators to rank each pair in comparison to 50 randomly selected pairs. This is a reasonable compromise, but it still results in a daunting annotation task, and makes the quality of the dataset depend on a random selection of comparisons.The problem of reliability is addressed by Luong et al. (2013) which included many rare words in their dataset, and thus allowed an annotator to indicate "Don't know" for a pair if they does not know one of the words. The problem with applying this approach as a more general reliability indicator is that the annotator confidence level is subjective and not absolute.

Kiritchenko and Mohammad (2017) present an analysis about Best–worst scaling (BWS) (Louviere 1991) as a method for data annotation as an alternative in order to overcome the problems of the Rating Scales (RS) method. In fact, RS is a widely used method for data annotation; however, it suffers from the difficulty in maintaining inter- and intra-annotator consistency. They show that with the same total number of annotations, BWS produces significantly more reliable results and high-quality annotations than the rating scale.

3.2.3 Annotation related task

The instructions imposed to the annotators should be so clarified in order to not overlap between the semantic similarity or relatedness tasks. For example, the antonym relation should be indicated as high relatedness rating or low one because its treatment differs between several SR-related datasets.

Agirre et al. (2009) made on the WordSim353 (Finkelstein et al. 2002) designed as word semantic relatedness dataset, showed that there is several word pairs evaluated according

[5] The term preferred-relation (such as hyponym-hypernym pairs) is used to denote the relation which the model should prefer, and unpreferred-relation to denote any other relation.

to their similarity which leads to the extraction of the dataset AG203. In contrast to gold standards such as WordSim353 and MEN3000, SimLex999 explicitly quantifies similarity rather than association or relatedness so that pairs of entities that are associated but not actually similar (e.g. *Freud/psychology*) have a low rating. Also, there is the problem of antonyms which are judged with high relatedness rating in RG65 and WordSim353.

3.2.4 Quality control

As there is no agreed gold standard for this kind of task (recall that the dataset annotation process is a way to establish this kind of gold standard), quality control is a hard task. If a human annotator disagrees with the majority, this might be due to fraudulent behavior or just a different valid opinion. However, it is possible to filter out annotators that provide the same value for each word pair or other implausible behavior. This kind of result is more likely in a crowd-sourcing setting than when there is more control over the participants like in a lab study. For example Radinsky et al. (2011) use pairs from an already established dataset (WS353) to discard poor-quality crowdworkers. Some researchers also use duplicate word pairs (e.g. Gerz et al. 2016; Halawi et al. 2012) in order to detect unreliable annotations. However, it remains an open question what level of variance on duplicate pairs is still acceptable. To verify the agreement between raters in MTurk771 (Halawi et al. 2012), they randomly split the raters into two groups, each including at least 10 Mechanical Turk workers. They then averaged the numeric judgments for each word pair among the raters in each of the two sets, thus yielding a vector of average judgments for each set. Finally, they computed the correlation between the mean judgment vectors of the two sets.

3.3 Discussion

Table 3 listed the monolingual datasets designed for more several languages in the context of semantic similarity/relatedness measures. In the context of the multilingual similarity, Table 4 shows the translation of some known datasets, initially proposed for the English language, to several other languages. Table 5 shows the cross-lingual datasets in the context of SemEval 2017 (Camacho-Collados et al. 2017).

Despite the numerous datasets and evaluation approaches presented in this study, evaluating semantic relatedness measures is both conceptually and practically an open challenge. Many biases are yet to be excluded from future datasets to improve evaluation process. For instance we propose that:

- Proposing a process including the different steps for the creation of the datasets. In fact, a great diversification exists in the research works focusing on the dataset construction.
- Following the best-worst scaling for the annotation step since it shows more reliability than rating scales.
- Providing measures for the quality quantification of the datasets. This evaluation can be based also on the scores of the semantic similarity measures through removing the dataset outliers.
- Translated datasets should follow an other annotation process via native speakers. In fact, a polysemous word does not express the most frequent sense in each language.
- Fixing a criteria set allowing the acceptation of a proposed dataset as an evaluation benchmark into the research community.

Table 3 Overview of datasets

Name	Year	# Pairs	POS	Scores	Language	Domain	Task	References
RG65	1965	65	N	[0, 4]	en	G	sim	Rubenstein and Goodenough (1965)
MC30	1991	30	N	[0, 4]	en	G	sim	Miller and Charles (1991)
RD37	2000	37	V	[0, 5]	en	G	sim	Resnik and Diab (2000)
WS353	2002	353	N, V, A	[0, 10]	en	G	rel	Finkelstein et al. (2002)
MeSH2	2005	36	N	[0, 1]	en	B	sim	Hliaoutakis (2005)
Gur350	2006	350	N, V, A	[0, 4]	de	G	rel	Gurevych (2006)
YP130	2006	130	V	[0, 4]	en	G	sim	Yang and Powers (2006)
ZG222	2006	222	N, V, A	[0, 4]	de	G	rel	Torsten and Iryna (2006)
Zie55	2006	55	NE	[0, 5]	en	G	rel	Ziegler et al. (2006)
MiniMayoSRS	2007	30	N	[0, 4]	en	B	sim	Pedersen et al. (2007)
MC30-rel	2008	30	N	[0, 4]	en	G	rel	Gracia and Mena (2008)
AG203	2009	203	N, V, A	[0, 10]	en	G	sim	Agirre et al. (2009)
UMNSRS-Rel	2010	587	N	NN	en	B	rel	Pakhomov et al. (2010)
UMNSRS-Sim	2010	566	N	NN	en	B	sim	Pakhomov et al. (2010)
MT287	2011	287	N, V, A	[0, 5]	en	G	rel	Radinsky et al. (2011)
MayoSRS	2011	101	N	[0, 10]	en	B	rel	Pakhomov et al. (2011)
SCWS	2012	1762	N	[0, 10]	en	G	rel	Huang (2012)
MTurk771	2012	771	N, V, A	[0, 5]	en	G	rel	Halawi et al. (2012)
ReWord26	2012	26	NE		en		rel	Pirrò (2012)
Atlasify240	2012	240	NE	{0, 4}	en	G	rel	Hecht et al. (2012)
MEN3000	2012	3000	N, V, A	0, 50	en	G	rel	Bruni et al. (2014)
Words	2012	240	N	[0, 10]	cn	B	rel	Wang et al. (2011)
WP300	2013	300	N	[0, 5]	en	G	sim	Li et al. (2013)
Alm70	2013	70	N	{0, 4}	ar	G	sim	Almarsoomi et al. (2013)
SaifAr	2013	40	N	{0, 4}	ar	G	rel	Saif et al. (2014)
RW2034	2013	2034	N	[0, 10]	en	G	rel	Luong et al. (2013)
Rel122	2013	122	N	[0, 4]	en	G	rel	Szumlanski et al. (2013)

Table 3 (continued)

Name	Year	# Pairs	POS	Scores	Language	Domain	Task	References
WP300	2013	300	MWE	[0, 1]	en	G	sim	Li et al. (2013)
MartinezAldana	2013	28	N	[0, 1]	en	G	sim	Gil and Montes (2013)
SL7576	2014	7576	N	[1, 5]	en	G	sim	Silberer and Lapata (2014)
SimLex999	2014	999	N, V, A	[0, 10]	en	G	sim	Hill et al. (2015)
Bekar143	2014	143	V	[0, 1]	en	G	sim	Baker et al. (2014)
GeReSiD50-sim	2014	50	N	[0, 1]	en	Geo	sim	Ballatore et al. (2014)
GeReSiD50-rel	2014	50	N	[0, 1]	en	Geo	rel	Ballatore et al. (2014)
Ugur	2016	101	N, V, A	[0, 5]	tr	G	rel	Sopaoglu and Ercan (2016)
PKU	2016	500	N, V, A	[1, 10]	cn	G	sim	Wu and Li (2016)
SimVerb3500	2016	3500	V	[0, 10]	en	G	sim	Gerz et al. (2016)
Gujarati-WS	2017	163	N, V, A	[0, 10]	gu	G	sim	Akhtar et al. (2017)
Punjabi-WS	2017	143	N, V, A	[1, 10]	pa	G	sim	Akhtar et al. (2017)
Tamil-WS	2017	97	N, V, A	[0, 10]	ta	G	sim	Akhtar et al. (2017)
Telugu-WS	2017	111	N, V, A	[0, 10]	te	G	sim	Akhtar et al. (2017)
Urdu-WS	2017	100	N, V, A	[0, 10]	ur	G	sim	Akhtar et al. (2017)
MJ	2017	12886	N, V, A	[0, 1]	ru	G	sim	Panchenko et al. (2017)
GTRD	2018	66	N	[0, 1]	en	Geo	rel	Chen et al. (2018)
ViData	2018	400	N	[0, 10]	vi	G	sim	Nguyen et al. (2018)
JWSDNoun	2018	1103	N	[0, 10]	ja	G	sim	Sakaizawa and Komachi (2017)
JWSDVerb	2018	1464	V	[0, 10]	ja	G	sim	Sakaizawa and Komachi (2017)
JWSDAdjective	2018	960	Adj	[0, 10]	ja	G	sim	Sakaizawa and Komachi (2017)
JWSDAdverb	2018	902	Adv	[0, 10]	ja	G	sim	Sakaizawa and Komachi (2017)
AnlamVer	2018	500	N, V, A	[0, 10]	tr	G	sim	Ercan and Yildiz (2018)
WiC	2019	5428	N, V	NA	en	G	sim	Pilehvar and Camacho-Collados (2019)

Version as used by Luong et al. (2013) excluding identical surface forms in different senses

en English, *ar* Arabic, *tr* Turkish, *ja* Japanese, *cn* Chinese, *gu* Gujarati, *pa* Punjabi, *ta* Tamil, *te* Telugu, *ru* Russe, *ur* Urdu, *POS* part of speech), *N* Noun, *V* Verb, *A* adjective and adverb, *NE* named entities, *MWE* multiword expression, *G* general, *Geo* geographic, *B* biomedical, *sim* similarity, *rel* relatedness, *NA* not available

Table 4 Translations of datasets

Language	Year	References
(a) Translations of RG65		
Arabic	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Chinese	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Dutch	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Farsi	2015, 2016, 2018	Camacho-Collados et al. (2015), Freitas et al. (2016), Barzegar et al. (2018)
French	2011, 2016, 2018	Joubarne and Inkpen (2011), Freitas et al. (2016), Barzegar et al. (2018)
German	2005, 2016, 2018	Gurevych (2005), Freitas et al. (2016), Barzegar et al. (2018)
Italian	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Portuguese	2014, 2016, 2018	Granada et al. (2014), Freitas et al. (2016), Barzegar et al. (2018)
Punjabi	2017	Akhtar et al. (2017)
Russian	2016, 2018	Panchenko et al. (2016), Freitas et al. (2016), Barzegar et al. (2018)
Spanish	2015, 2016, 2018	Camacho-Collados et al. (2015), Barzegar et al. (2018), Freitas et al. (2016)
Swedish	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Hungarian	2013	Tóth (2013)
Czech	2017	Konopik et al. (2017)
(b) Translations of WS353		
Arabic	2009, 2018	Hassan et al. (2009), Barzegar et al. (2018)
Chinese	2012, 2018	Jin and Wu (2012), Barzegar et al. (2018)
Czech	2016, 2017	Cinková (2016), Konopik et al. (2017)
Dutch	2014	Postma and Vossen (2014), Barzegar et al. (2018)
German	2015, 2018	Leviant and Reichart (2015), Barzegar et al. (2018)
Italian	2015, 2018	Leviant and Reichart (2015), Barzegar et al. (2018)
Romanian	2009	Hassan et al. (2009)
Spanish	2009, 2018	Hassan et al. (2009), Barzegar et al. (2018)
Arabic	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Chinese	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Dutch	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)

Table 4 (continued)

Language	Year	References
Farsi	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
French	2011, 2018	Freitas et al. (2016), Barzegar et al. (2018)
German	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Italian	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Portuguese	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Russian	2015, 2016, 2018	Leviant and Reichart (2015), Freitas et al. (2016), Panchenko et al. (2016), Barzegar et al. (2018)
Spanish	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Swedish	2016	Freitas et al. (2016), Barzegar et al. (2018)
Thai	2019	Netisopakul et al. (2019)
(c) Translations of SimLex999		
German	2015, 2018	Leviant and Reichart (2015), Barzegar et al. (2018)
Italian	2015, 2018	Leviant and Reichart (2015), Barzegar et al. (2018)
Russian	2015	Leviant and Reichart (2015)
Vietnamese	2017	Tan et al. (2017)
Dutch	2016	Freitas et al. (2016)
French	2011, 2018	Freitas et al. (2016), Barzegar et al. (2018)
German	2016	Freitas et al. (2016)
Italian	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Portuguese	2016	Freitas et al. (2016)
Spanish	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Swedish	2016, 2018	Freitas et al. (2016), Barzegar et al. (2018)
Farsi	2018	Barzegar et al. (2018)
Thai	2019	Netisopakul et al. (2019)

Table 5 Number of word pairs in each dataset proposed for the multilingual semantic similarity sub-task in SemEval 2017 (Camacho-Collados et al. 2017)

	EN	DE	ES	IT	FA
EN	500	914	978	970	952
DE		500	956	912	888
ES			500	967	967
IT				500	916
FA					500

4 Intrinsic evaluation: process and metrics

In an intrinsic evaluation, the results of a relatedness measure are directly compared to a gold standard dataset.

4.1 Inter-annotator agreement

It is common practice to compare the annotations of multiple people using inter-annotator agreement (Artstein 2017). DKPro agreement (Meyer et al. 2014) is an open source Java library for measuring inter-annotator agreement. The computed inter-annotator agreement can also serve as an upper limit for the performance of a semantic relatedness measure on the same data (Resnik and Lin 2010; Zesch and Gurevych 2010). Inter-annotator agreement has also been used for analyzing dataset, e.g. Zesch and Gurevych (2010) found largely varying agreement scores for the two subsets within the WS353 dataset and thus suggested treating them separately in evaluation. Note that the subsets appear in the literature as Fin153/Fin200 or WS153/WS200.

In contrast to *inter*-annotator agreement, the *intra*-annotator agreement measures the agreement of a judge with herself over time. Unfortunately, only few experiments with intra-annotator agreement have been performed in previous work.

4.2 Evaluation metrics

The three most widely used metrics for evaluating relatedness measures are Pearson correlation, Spearman correlation and Kendall tau, the latter two being generally suggested for data not normally distributed. They indicates how well the results of a measure resemble human judgments, where a value of 0 means no correlation and 1 means perfect correlation.

Pearson (r) The Pearson product-moment correlation coefficient r is calculated as:

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2)(\sum x_i)^2}\sqrt{n(\sum y_i^2)(\sum y_i)^2}} \tag{1}$$

where x_i refers to the ith element in the list of human judgments, y_i refers to the corresponding ith element in the list of computed values, and n refers to the number of word pairs.

Spearman (ρ) Is used to correlate word pair rankings. In case a measure returns scores instead of rankings, the ordered scores can be easily converted into ranks. Spearman correlation is computed as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

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

Harmonic mean (μ) An ideal system should be able to demonstrate a linear relationship with the human judgment and also to satisfy the relative order imposed by those judgments. In a sense, a good system should maintain the correct ranking between word pairs and, at the same time, correctly quantify the strength of the relatedness for a given word pair. Therefore, the harmonic mean (μ) of the Pearson and Spearman correlation coefficients (Hassan et al. 2012), is reported in as follows:

$$\mu = \frac{2r \times \rho}{r + \rho} \tag{3}$$

This measure has e.g. been used as the official evaluation metric in the "SemEval-2017 Task 2, Multilingual and Cross-lingual Semantic Word Similarity" (Camacho-Collados et al. 2017).

Kendall's tau (τ) Is even less sensitive to outliers than Spearman and is computed as follows:

$$\tau = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_i - x_j) \times sgn(y_i - y_j)}{n(n-1)} \tag{4}$$

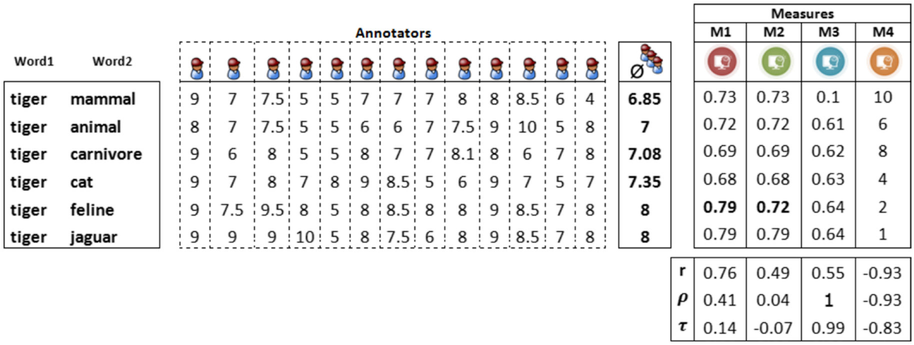
$$sgn(x_i - x_j) = \begin{cases} -1 \; if \; (x_i - x_j < 0) \\ 1 \; if \; (x_i - x_j) > 0 \\ 0 \; if \; (x_i - x_j) = 0 \end{cases} \tag{5}$$

$$sgn(y_i - y_j) = \begin{cases} -1 \; if \; (y_i - y_j < 0) \\ 1 \; if \; (y_i - y_j) > 0 \\ 0 \; if \; (y_i - y_j) = 0 \end{cases} \tag{6}$$

The product of the sign functions $sgn(x_i - x_j) \times sgn(y_i - y_j)$ can be interpreted as a concordance indicator. It is equal to 1 for matching pairs and -1 for discordant pairs.

In the case when there are no tied values in the data, Kendall's coefficient produces consistently narrower confidence intervals, and might thus be preferred on that basis. However, if there are any ties in the data, irrespective of whether the percentage of ties is small or large, Spearman's measure returns values closer to the desired coverage rates, whereas Kendall's results differ more and more from the desired level as the number of ties increases, especially for large correlation values.

Example In Fig. 3, we give an example for computing the evaluation metrics and show that they can be quite sensitive to small modification of relatedness results. For example, measures M1 and M2 return identical scores except for one word pair where their difference is rather small (0.79 vs. 0.72). However, this results in rather dramatic changes in the evaluation ($r = 0.76 \rightarrow r = 0.49$ and $\rho = 0.41 \rightarrow \rho = 0.04$). Moreover, Fig. 3 explains the complementarity between the correlation coefficients based on values (Pearson) and those based on the ranks (Spearman and Kendall).In fact, the automatic provided values by M1 and the M3 give a significant difference between the computed correlations (Pearson and



		Annotators														Measures			
Word1	Word2														Ø	M1	M2	M3	M4
tiger	mammal	9	7	7.5	5	5	7	7	7	8	8	8.5	6	4	**6.85**	0.73	0.73	0.1	10
tiger	animal	8	7	7.5	5	5	6	6	7	7.5	9	10	5	8	**7**	0.72	0.72	0.61	6
tiger	carnivore	9	6	8	5	5	8	7	7	8.1	8	6	7	8	**7.08**	0.69	0.69	0.62	8
tiger	cat	9	7	8	7	8	9	8.5	5	6	9	7	5	7	**7.35**	0.68	0.68	0.63	4
tiger	feline	9	7.5	9.5	8	5	8	8.5	8	8	9	8.5	7	8	**8**	**0.79**	**0.72**	0.64	2
tiger	jaguar	9	9	9	10	5	8	7.5	6	8	9	8.5	7	8	**8**	0.79	0.79	0.64	1

	M1	M2	M3	M4
r	0.76	0.49	0.55	-0.93
ρ	0.41	0.04	1	-0.93
τ	0.14	-0.07	0.99	-0.83

Fig. 3 Overview of intrinsic evaluation based on a subset of the WordSim353 dataset. Metrics used: Pearson (r), Spearman (ρ) and Kendall (τ). The dashed table concerns the annotations leading to the final human judgments (the average) exploited for evaluating the measures performance

Spearman). So, it is important to express the performance of a proposed measure using the different coefficients because an efficient measure should ensure the proportional values to those attributed by annotators and keeping the semantic distance order gave by them. The M4 shows the fact that the provided similarity values can be inversely proportional (using the measures based on the counting edges between concepts in a knowledge graph) to the experts judgments which leads to negative correlations that express good performance when they getting closer to -1.

4.3 Scatterplots

As all metrics can be misleading, it is good practice to also visually inspect the results. As it is illustrated in Fig. 4, a scatterplot can be used for this purpose, where the gold standard values are plotted on the x-axis and the corresponding values returned by the relatedness measure on the y-axis. A perfect correlation would lead to a straight line through (0, 0) and (1, 1). Especially outliers influencing the results can often be easily spotted in that way.

4.4 Coverage

A factor influencing the evaluation metrics is *coverage*. It is defined as the percentage of word pairs in a dataset for which a certain measure does not return a defined value (Ben Aouicha et al. 2016b). Note that some relatedness measures do not return undefined values, but either zero or some default value in the middle of the score distribution. Thus coverage can usually not be reliably computed and relatedness measures should be compared on the full dataset.

5 Semi-intrinsic evaluation

There are tasks that are somewhat in between intrinsic and extrinsic and that we call *semi-intrinsic*. In the intrinsic evaluation (as described above), evaluation datasets are used that contain judgments of semantic similarity or relatedness so that the performance of the

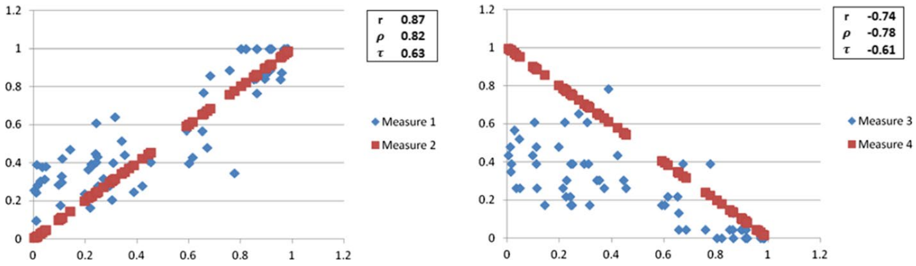


Fig. 4 Scatterplot for the correlations between the Human judgments of the dataset RG65 (Rubenstein and Goodenough 1965) (x-axis) and the values provided by semantic similarity measures (y-axis). The red shapes present the best measure behavior. The correlations are: Pearson (r), Spearman (ρ) and Kendall (τ). The right figure presents an example of a measure conversely propositional to the Human judgments

semantic relatedness measure can be directly measured. In the extrinsic evaluation, semantic relatedness measures are only used as one component in a larger system and performance is measured in terms of the host task, e.g. as precision/recall in a retrieval setting or user satisfaction. Semi-intrinsic tasks fall between both classes, as they use small, manually-annotated evaluation datasets and results can usually be directly attributed to the performance of the underlying semantic relatedness approaches. Semi-intrinsic tasks are also not end-user centric tasks, but often only build for evaluation purposes. Semi-intrinsic tasks do not require often the kind of imprecise manual annotation of semantic relatedness values.

5.1 Text similarity

The research area that is closest in spirit to word similarity and relatedness is sentence/text similarity. It deals with the similarity between long passages such as sentences and paragraphs. As shown in Fig. 5, semantic relatedness measures are used to calculate a match between the terms of two texts and the scores are then aggregated (Mihalcea et al. 2006).

For the sentence level, Ben Aouicha et al. (2015) presented a new method for computing sentence semantic similarity by exploiting a set of its characteristics, namely Features-based Measure of Sentences Semantic Similarity (FM3S). There are several datasets designed for the evaluation of text semantic similarity (Nguyen et al. 2019).

Additionally, it has been argued (Bär et al. 2011) that text similarity involves text dimensions beyond content such as structure and style that are not easily captured by semantic relatedness measures. Nonetheless, in the first SemEval text similarity challenge in 2012 (Agirre et al. 2012), the best results was achieved by a combination of many semantic relatedness measures (Bär et al. 2012a).

5.2 Word choice

It is a task that concerns the identification from four candidate words the one that is closest in meaning to a given source word. Clearly a task with relatively little practical relevance, but a quite direct way of evaluating semantic relatedness measures. These types of questions are often found in many English essays such as TOEFL (Test of English as a Foreign Language). This task has been adopted by a number of studies including (Jarmasz and Szpakowicz 2003; Ruiz-Casado et al. 2005; Tsatsaronis et al. 2010a; Hadj Taieb et al. 2013).

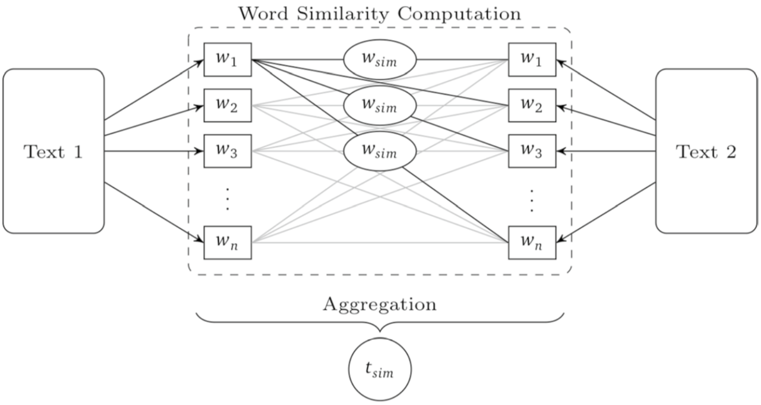


Fig. 5 Text similarity approaches based on word semantic similarity/relatedness Figure adapted from Bär et al. (2015)

5.3 Meaning-based segmentation

Here, similarity measures are used to cluster words based on classes or groups already fixed. Bollegala et al. (2007) evaluated the performance of their proposed measure in capturing the semantic similarity between named-entities by setting up a community mining task. They select 50 personal names from 5 communities from the Open directory. Project[6]. Cilibrasi and Vitanyi (2007) evaluated their method by the hierarchical grouping of a manually created set of words. In the biomedical field, Wang et al. (2007) grouped genetic products based on the calculation of the relatedness between Gene Ontology terms used to annotate the functions related to each gene product.

5.4 Word analogy

Word Analogy[7] concerns for determining the proportional analogy holds between two word pairs: a:a* : : b:b* (a is to a* as b is to b*). For example, *Tokyo* is to *Japan* as *Paris* is to *France*. Wang et al. (2019a) compare semantic relatedness measures based on word embedding models based on two datasets adopted for word analogy evaluation task (Google dataset (Mikolov et al. 2013a) and MSR dataset (Mikolov et al. 2013b)).

5.5 Concept categorization

Here, the goal is to split a given set of words into different categorical subsets of words. For example, given the task of separating words into two categories, the model should be able to categorize words *sandwich*, *tea*, *pasta*, *water* into two groups. Three datasets are known in concept categorization evaluation: (1) AP dataset (Almuhareb 2006), (2) the BLESS dataset (Baroni and Lenci 2011) and (3) the BM dataset (Baroni et al. 2010).The

[6] http://odp.org/.

[7] https://aclweb.org/aclwiki/Google_analogy_test_set_(State_of_the_art).

AP dataset contains 402 words that are divided into 21 categories. The BM dataset is a larger one with 5321 words divided into 56 categories. Finally, the BLESS dataset consists of 200 words divided into 27 semantic classes. Mandera et al. (2017) exploited this task combined the semantic measures for explaining human performance in psycholinguistic tasks.

5.6 Outlier detection

The goal is to find words that do not belong to a given group of words. This evaluator tests the semantic coherence of vector space models, where semantic clusters can be first identified. There are two datasets for the outlier detection task: (1) the WordSim-500 and (2) the 8–8–8 datasets. The WordSim-500 consists of 500 clusters, where each cluster is represented by a set of 8 words with 5–7 outliers (Blair et al. 2017). The 8–8–8 dataset has 8 clusters, where each cluster is represented by a set of 8 words with 8 outliers (Camacho-Collados and Navigli 2016). Santus et al. (2018) exploited this task for evaluating SR and SS measures.

6 Extrinsic evaluation: application-based

In contrast to the direct intrinsic evaluation, extrinsic evaluation is more indirect as semantic relatedness measures are used in another application and the performance of this application is evaluated. We do not cover evaluation protocols here, as they depends on the application, but give a short summary of the most important application areas and how semantic relatedness is used therein.

6.1 Search and retrieval

Semantic relatedness measures are often applied in information retrieval (Lee et al. 1993) to overcome the shortcomings of the classical bag-of-words model. For example, a document containing the term *bus* is not a match for the query *public transport* under the bag-of-words model. However, both terms deal with the same theme “means of transport”. Semantic relatedness measures can be used to bridge this vocabulary gap between query and document either directly as a retrieval model (Ensan and Du 2018) (i.e. the relevance of a document is directly computed based on a semantic relatedness measure), as well as when being used for query expansion (Sahami and Heilman 2006; Egozi et al. 2008) or for re-ordering the retrieval results. Those methods have been applied in many domains including searching for proteins (Lord et al. 2003), video sequences (Schickel-Zuber and Faltings 2007), medical documents (Angelos et al. 2006), or job search (Gurevych et al. 2007). When having established the most relevant documents, users, or resources based on semantic relatedness measures, it is a logical step to also use that within *recommender systems*, e.g. blog recommendation (Li and Chen 2009) or user recommendation (Meo et al. 2011).

6.2 Text summarization and segmentation

TextRank (Mihalcea and Tarau 2004) is a well-known text summarization algorithm based on semantic relatedness measures. It computes the importance of words and sentences within a text by computing their pairwise relatedness and then measuring the centrality of concepts or sentences in the resulting graph. Similar techniques have been used for speech (Gurevych and Strube 2004) and meeting summarization (Xie and Liu 2008). Moreover, the SR measures are used for resolving the keyphrase extraction task (Zesch 2010; Xie et al. 2010).

The computation of pairwise relatedness between sentences is also in the core of many *text segmentation* algorithms. It can either work on a fully connected graph like GraphSeg (Glavas et al. 2016) or only between consecutive sentences where thematic breaks can be identified as gaps within the relatedness scores (Kozima 1993).

6.3 Relation extraction

A rather direct application of semantic relatedness measures is *automatic thesauri generation*. We simply need to measure the strength of the relationship between a term and all other terms in the vocabulary and then use the ones with the strongest relationship as candidates for the thesaurus (Curran 2002; Panchenko and Morozova 2012). Similar methods are also used for ontology alignment, where for a term in one ontology we are looking for the best matching term in another ontology based on semantic relatedness measures (Gracia and Mena 2008; Lin and Sandkuhl 2008). When combined with a method to establish the type of relationship between two terms, the method can also be used to identify the type of relationship, e.g. synonymy, hyponymy, or meronymy (Marie-Francine 2013). Semantic relatedness measures can also be used to establish domain specific relations such as protein–protein interactions (Zhang and Tang 2016). If we consider the special case when the terms are actually named entities, the same methodology can also be used for *Entity Linking* (Sánchez et al. 2011) and co-reference resolution (Ponzetto and Strube 2006; Yang and Su 2007). In the context of multidimensional schemas used to model data warehouse systems, Salem and Ben-Abdallah (2015) proposed an approach integrating the semantic similarity computation to solve the problem in relation with the difference between the names used in the design of the star schema and the names existing in the data source schema.

6.4 Word sense disambiguation

Semantic relatedness measures have been in the core of Word Sense Disambiguation (WSD) for quite a while [a comprehensive overview is given in Navigli (2009), Wang et al. (2019b)]. Most approaches select a word sense by maximizing the semantic relatedness between a context window around the ambiguous word and a representation of the word sense [e.g. glosses (Banerjee and Pedersen 2003; Patwardhan et al. 2006; Gracia and Mena 2008) or the taxonomic structure (Ben Aouicha et al. 2018b)]. The advancements in the semantic relatedness, as a research field, for other languages and for specific domains, lead to an improvement for treating the WSD for these languages as Alkhatlan et al. (2018) in Arabic or Gabsi et al. (2017) for the biomedical domain.WSD includes the specific task Named Entity Disambiguation which exploits the SS measures (Zhu and Iglesias 2018). WSD researchers community has an evaluation campaign exploiting several benchmarks.

6.5 Spelling correction

While out-of-vocabulary errors are easy to detect, it remains a challenging problem to detect real-word spelling errors (also called *malapropisms*) (Hirst and Budanitsky 2005). Semantic relatedness measures are applied within this context to determine whether a word does not fit its semantic context, i.e. the semantic relatedness with all other words in its environment is low and there is another word that would fit in much better (Budanitsky and Hirst 2006; Zesch 2012).

6.6 Plagiarism detection

While plagiarism in the form of verbatim lifting of text from some source document can be quite easily detected based on string matching, plagiarists often disguise their actions by substituting words with closely related words like synonyms. Computational methods can counter this effect with the help of semantic relatedness measures by computing the similarity between the text under consideration and the source texts (Abdi et al. 2015; Bär et al. 2012b; Tsatsaronis et al. 2010a, b). As there usually is a very large number of potential source texts, a cheaper method (e.g. text fingerprints) are used to compile a small set of probably plagiarism cases and the expensive computation based on semantic relatedness measures is only applied to this smaller set.

6.7 Pervasive computing

In pervasive computing, the semantic similarity measure is a tool to allow services to be chosen and classified according to their relevance to a given query, and a user's profile and preferences to be compared to those of other users in order to recommend similar services. Finally, semantic similarity aims to evaluate the similarity between application components in order to propose the most relevant one in a current context. Pervasive computing system includes a context-aware applications that interact with the physical environment and the user's system in order to provide appropriate services. In such research field, semantic similarity measures have been applied at several levels [see Fig. 2 in Guessoum et al. (2015)] such as the service discovery by the matching the description of a request with available services. Guessoum et al. (2015, Table 3) gave an overview about the studies on semantic similarities in pervasive computing.

6.8 Sentiment analysis

Sentiment analysis or opinion mining is a domain that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from a written language; it had become a very active area of scientific research in recent years, especially with the development of social networks like Facebook and Twitter. The evaluation of the effectiveness of the proposed models is ensured through the availability of public datasets (Araque et al. 2019). Madani et al. (2019) proposed two methods to classify the tweets into three classes: positive, negative and neutral (subjectivity classification). These methods are both based on the calculation of semantic similarity (using WordNet) between the collected tweets and the present classes.

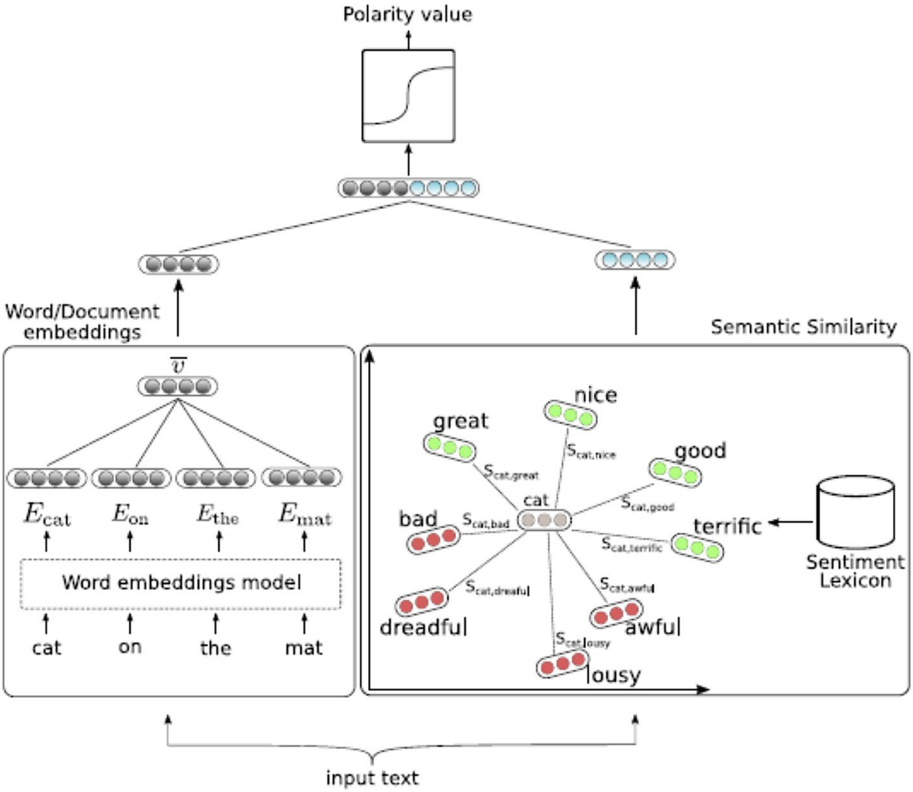


Fig. 6 System architecture diagram for the sentiment analysis based on knowledge-based and word embedding-based similarity measures (Araque et al. 2019)

Araque et al. (2019) proposed a novel method of utilizing sentiment lexicons, which is based on a semantic similarity metric between text words and lexicon vocabulary as it is described in Fig. 6. Their proposal consists of a sentiment analysis that uses this lexicon-based semantic similarity as features, as well as embedding-based representations.

The natural advancement of this research field leads to the apparition of the Aspect Based Sentiment Analysis (ABSA). It focuses on understanding user opinion about different aspects of products, services or policies that can be used for improving and innovating in an effective way. Araque et al. (2016) proposed a hybrid model for ABSA consisting of a word embeddings model used in conjunction with semantic similarity measures in order to develop an aspect classifier module.

7 Tools and shared tasks

This section provides an overview of existing software solutions and source code libraries dedicated to semantic similarity/relatedness computation and analysis.

7.1 Replicability and reproducibility challenges

Cohen et al. (2018) defined the replicability or repeatability as a property of an experiment: the ability to repeat—or not—the experiment described in a study. As for the reproducibility is a property of the outcomes of an experiment: arriving—or not—at the same conclusions, findings, or values. Wieling et al. (2018) state that basing on a reproducibility study in the field of NLP only a third part of the published works (36.2%) provided their source code; therefore, researchers found several problems for being reproduced exactly.

Providing implementations of existing similarity measures to end- users and therefore contributing to the large adoption of semantic measures are not a little thing considering that implementations are most often very technical and have to go through the extraction of the needed semantic information from the semantic resources. Second, by providing common platforms for measure evaluation, these development projects are also essential to support research contributions in the field.

Lastra-Díaz et al. (2017) presented their work for remedying the lack of a set of self-contained and easily reproducible experiments that allow the research community to be able to replicate methods and results reported in the literature exactly, even without the need for software coding. The lack of reproducible experiments, together with the aforementioned lack of software libraries covering the most recent methods, and the difficulties in replicating methods and experiments exactly have contributed, with few exceptions, to improvable reproducibility practices in the area. Many works introducing similarity measures during the last decade have only implemented or evaluated classic measures, avoiding the replication of similarity measures introduced by other researchers.

Ordering of tools is therefore made considering subjective criteria such as popularity of the tool in the community, functionalities provided, source code base, documentation available, support. The large majority of projects focus on corpus-based or knowledge-based measures.

7.2 Tools for corpus-based semantic measures

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora.

- *DKPro Similarity*[8] is a framework dedicated to the comparison of pairs of words and pairs of texts (Bär et al. 2013). It provides numerous implementations of state-of-the-art semantic relatedness measures in Java—some knowledge-based measures are also available to compare WordNet synsets. This project is part of the DKPro Core[9] project which develops a collection of software components dedicated to NLP. These components have the interesting properties to be based on the Apache UIMA (Unstructured Information Management Applications) framework. The last version of DKPro Similarity has been release in April 2018 (version 2.3.0), it is distributed under the open-source Apache Software License.

[8] https://github.com/dkpro/dkpro-similarity/releases.

[9] https://dkpro.github.io/dkpro-core/.

- *Semilar*[10] proposes a software and development environment dedicated to corpus-based semantic measures (Rus et al. 2013). It provides both a Java library and a GUI-based interface. Semilar can be used to compare words and sentences. Numerous measures have been implemented, and several distributional models are made available.
- *Disco*[11] is an open-source Java library dedicated to the semantic similarity computation between words (Kolb 2008). The tool is distributed under the Apache License, version 2.0. Several measures are implemented. Interestingly, numerous languages are also supported: Arabic, Czech, Dutch, English, French, German, Italian, Russian, and Spanish. Last version available to date is version 3.0 (released in June 2018).
- *NLTK*[12] is a NLP platform developed in Python (Bird 2006). It contains a module dedicated to WordNet which provides specific semantic measures for comparing two synsets. NLTK is distributed under the Apache 2.0 license.
- *GenSim*[13] is a Python platform dedicated to statistical semantics (Řehůřek and Sojka 2010). It is well documented and provides several efficient distributional models and measures implementations. Gensim is distributed under the LGPL Licence. Both open-source and business supports are provided. The last release has been developed in January 2019.
- *WikiBrain*[14] is a documented Java project that proposes Wikipedia-based algorithms for semantic relatedness computation (Sen et al. 2014). It can be used to compare both sentences and Wikipedia pages (topics). Efforts have been made to ensure efficiency and shorten computational time. WikiBrain is distributed under the Apache 2.0 license.
- *TakeLab*[15] is a semantic text similarity system in Python code that can be used to compare two sentences (Šarić et al. 2012). It is licensed under a derivative of a BSD-license that requires proper attribution. This source code corresponds to a submission proposed to the SemEval evaluation campaign.
- *SemSim*[16] is a Java library that can be used to evaluate the semantic relatedness of words and to compute distributional models from texts. It can also be used to compute distributional models. The source code is distributed under license LGPL v3. The last version has been released in 2013.
- *Mechaglot*[17] can be used to compare sentences. It is distributed under the Creative Commons Attribution-ShareAlike 4.0 International License. This project is still under development and to date only provides an alpha version for developers.

7.3 Tools for word embedding-based semantic measures

Word embedding is a popular semantic model which represents words and sentences in computational linguistics systems and machine learning models. In recent years a large set of algorithms for both generating and consuming word embedding models (WEMs)

[10] http://www.semanticsimilarity.org/.

[11] http://www.linguatools.de/disco/disco-download_en.html.

[12] http://www.nltk.org/.

[13] https://radimrehurek.com/gensim.

[14] https://shilad.github.io/wikibrain/.

[15] http://takelab.fer.hr/sts.

[16] http://www.marekrei.com/projects/semsim/.

[17] http://mechaglot.sourceforge.net.

have been proposed, which includes corpus pre-processing strategies, WEM algorithms or weighting schemes, vector compositions and distance measures (Lastra-Díaz et al. 2019b).

- *S-SPACE*[18] is a library to support the construction of count based distributional methods unifying different approaches in a common JAVA API (Jurgens and Stevens 2010).
- *DEEPLEARNING4J*[19] , on the other hand, is a Java library which concentrates predictive-based models. Its API contains methods to access word vectors and to find nearest neighbours (kNN).
- *GENSIM*[20] is one of the most popular word-embedding toolkits, mainly credited to its efficient implementation of nearest neighbours function (Řehůřek and Sojka 2010). GENSIM is written in python and apart from its kNN function, it supports the generation of predictive-based models and methods to access word vectors.
- *DISSECT*[21] (DIStributional SEmantics Composition Tookit) focuses on vector compositions (Dinu et al. 2013). DISSECT is a PYTHON library containing methods to generate vector representation of sentences from the vector of its constituting words. DISSECT partially supports the generation of count-based models and brings an integrated baseline framework for evaluation purposes.
- *JoBimText*[22] is a semantic similarity tool that implements its own algorithm named JoBim (Biemann and Riedl 2013). The tool supports the construction of the JoBim model and also calculates semantic relatedness of pairs of terms, finds nearest neighbours and offers a native web server.
- *EasyESA*[23] (Carvalho et al. 2014) and *DInfra* (Barzegar et al. 2015) are also two initiatives to deliver distributional semantics capabilities under a more specific set of distributional semantic models.
- *INDRA*[24] (Sales et al. 2018) describes a word embedding/distributional semantics framework which supports the creation, use and evaluation of word embedding models. INDRA provides a software infrastructure to facilitate the experimentation and customisation of multilingual WEMs, allowing end-users and applications to consume and operate over multiple word embedding spaces as a service or library. INDRA shares more than 65 pre-computed models in 14 languages.
- *HESML*[25] introduced by Lastra-Díaz et al. (2019b) and including word embedding models with the aim of providing a common software platform for the evaluation of both knowledge based and word embedding models.
- *DKPro*[26] is also extended to provide services ensuring of the use of the word embedding models in measuring of semantic relatedness (Horsmann and Zesch 2018).

[18] https://github.com/fozziethebeat/S-Space.

[19] https://deeplearning4j.org/docs/latest/deeplearning4j-nlp-word2vec.

[20] https://radimrehurek.com/gensim/.

[21] https://github.com/composes-toolkit/dissect.

[22] http://ltmaggie.informatik.uni-hamburg.de/jobimviz/.

[23] https://github.com/dscarvalho/easyesa.

[24] https://github.com/Lambda-3/Indra.

[25] https://data.mendeley.com/datasets/t87s78dg78/4.

[26] https://dkpro.github.io/dkpro-tc/.

7.4 Tools for knowledge-based semantic measures

There are a number of measures that were developed to quantify the degree to which two words are semantically related using information drawn from semantic networks (e.g. WordNet, MesH, SNOMED-CT).

- *SML*[27] stands for Semantic Measures Library (Harispe et al. 2014). This is a Java library dedicated to semantic measure computation, development and analysis. SML implements numerous measures for comparing concepts and groups of concepts defined into ontologies. In addition to the library, a command-line toolkit is also developed for non-developers. Both the library and the toolkit are optimized to handle large dataset and to ensure fast computation[28] . They are distributed under a GPL compatible license. Last release: 2017.
- *FastSemSim*[29] is a Python library dedicated to semantic similarity computation (Guzzi et al. 2012). It can be used to compare pairs of concepts and pairs of groups of concepts through the semantic similarity measures over ontologies. It provides implementations of several existing semantic similarity measures. FastSemSim is multiplatform, and it works with Python 2.x and 3 and it is Open Source, distributed under the GNU General Public License version 3. Last version: 2014.
- *SimPack*[30] is a Java library that can be used to compare pairs of concepts defined into ontologies (Bernstein et al. 2005). It implements a large variety of measures and can load RDF/OWL ontologies. SimPack is distributed under the LGPL license.
- *SemMF*[31] is a Java library that can be used to compare objects defined into RDF graphs. It is distributed under LGPL licence (Oldakowski and Bizer 2005).
- *OntoSim*[32] is a Java library dedicated to the comparison of ontologies (David and Euzenat 2008). Several measures provided in this library can also be used to compare objects defined into RDF graphs.
- *YTEX*[33] is a Java library that can be used to compare two concepts or two groups of concepts defined into an ontology (Garla and Brandt 2012). The library can be used with UMLS, SNOMED-CT and MeSH. It is also possible to load other ontologies through SQL queries.
- *Similarity Library*[34] is Java library that can be used to compare pairs of concepts defined into several ontologies (WordNet, MesH, GO) (Pirró and Euzenat 2010). The library is made available on request.
- *WordNet-Similarity*[35] is a Perl module dedicated to semantic similarity and relatedness measures between WordNet synsets (Pedersen et al. 2004). This package has been largely used in the literature. Last version: 2008.

[27] http://www.semantic-measures-library.org.

[28] A comparison with other tools is provided at: https://github.com/sharispe/sm-tools-evaluation.

[29] https://pypi.org/project/fastsemsim/.

[30] https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/.

[31] http://semmf.ag-nbi.de/doc/index.html.

[32] http://ontosim.gforge.inria.fr/.

[33] https://code.google.com/p/ytex/wiki/SemanticSim_V06.

[34] https://simlibrary.wordpress.com/.

[35] http://wn-similarity.sourceforge.net/.

- *WS4J*[36] is a Java library dedicated to the development of semantic measures for WordNet. It is distributed under the GPL licence. Last version: 2013
- *UMLS-Similarity*[37] is a Perl module that can be used to compare concepts defined into UMLS (McInnes et al. 2009).
- *OWLSim*[38] is a Java Library for the comparison of pairs of concepts defined in OWL format (Washington et al. 2009). Last version: 2017.
- *DOSim*[39] is an R package dedicated to semantic similarity computation for the Disease Ontology (Li et al. 2011). DOSim is distributed under the GPL licence. Last version: 2010.
- *DOSE*[40] is another R package dedicated to semantic similarity computation for the Disease Ontology. DOSE is distributed under the Artistic-2.0 licence.
- *Serelex*[41] is a "lexico-semantic search engine" (Panchenko et al. 2013). Given a query it returns a list of related words. This way allows discovering meaning of words in an interactive manner, search for related words and more.
- *Sematch*[42] is an integrated framework for the development, evaluation and application of semantic similarity for Knowledge Graphs. The framework provides a number of similarity tools and datasets, and allows users to compute semantic similarity scores of concepts, words, and entities Sematch[43] focuses on knowledge-based semantic similarity that relies on structural knowledge in a given taxonomy (Zhu and Iglesias 2017) (e.g. depth, path length, least common subsumer), and statistical information contents.
- HESML[44] is a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication datasets (Lastra-Díaz et al. 2017).
- WNetSS[45] is a Java API allowing the use of a wide WordNet-based semantic similarity measures pertaining to different categories including taxonomic-based, features-based and Information Content-based measures (Ben Aouicha et al. 2018a). It allows the extraction of the topological parameters from the WordNet "is a" taxonomy which are used to express the semantics of concepts (Hadj Taieb et al. 2014). Moreover, an evaluation module is proposed to assess the reproducibility of the measures accuracy that can be evaluated according to 10 widely used benchmarks through the correlations coefficients.

7.5 Discussion

As it is shown in this section, there are some tools provided for using semantic measures and supporting large scale comparisons of measures. For distributional measures, proposed software is not limited to a specific corpus of texts. So, they can therefore be used in a large

[36] http://code.google.com/p/ws4j/.

[37] http://umls-similarity.sourceforge.net/.

[38] https://github.com/monarch-initiative/owlsim-v3.

[39] http://210.46.85.150/platform/dosim/.

[40] http://www.bioconductor.org/packages/release/bioc/html/DOSE.html.

[41] http://serelex.cental.be/.

[42] http://sematch.cluster.gsi.dit.upm.es/.

[43] https://github.com/gsi-upm/sematch.

[44] https://github.com/jjlastra/HESML.

[45] http://wnetss-api.smr-team.org/.

diversity of contexts of use when the texts corpus is the semantic resource, e.g. DKPro[46] (Bär et al. 2013), Semilar[47] (Rus et al. 2013), Disco[48] (Kolb 2008) and swoogle[49] (Han et al. 2013). Conversely, software solutions dedicated to knowledge-based semantic measures are generally developed for a specific domain, e.g., WordNet (Pedersen et al. 2004) Ben Aouicha et al. (2018a), UMLS (McInnes et al. 2009). Large number of solutions are developed for the Gene Ontology such as InteGO2[50] which is a web tool for calculating the gene ontology (GO)-based gene semantic similarities using seven widely used GO-based similarity measurements (Peng et al. 2016). The diversity of software solutions is also beneficial as it generally stimulates the development of robust solutions.

As discussed in Harispe et al. (2015) and other contributions, the evaluation of semantic measures is mainly governed by empirical studies used to assess their accuracy according to expected scores/behaviours of the measures. Therefore, the lack of open-source software solutions implementing a large diversity of measures hampers their study. It explains, for instance, that evaluations of measures available in the literature only involve the comparison of a subset of measures which is not representative of the diversity of semantic measures available today. Initiatives aiming at developing robust open-source software solutions which give access to a large catalogue of measures must therefore be encouraged. It is worth noting the importance of these solutions being open-source.

The scientific community, related to the discussed research fields, lacks open-source software dedicated to the evaluation of semantic measures. Indeed, despite some initiatives such as DKPro Similar and CESSM[51] (Collaborative Evaluation of Semantic Similarity Measures) (Pesquita et al. 2009), evaluations are not made through a common framework as it is done in most communities. This is mandatory to finely compare and evaluate semantic measures in a large scale way. The development of such tools must ensure fair comparison of results, as well as experiment reproducibility.

Nevertheless, important efforts have recently been made to establish evaluation campaigns related to semantic measures such as the SemEval[52] conferences including several tasks. They have been organized in 2012–2019 in order to compare proposals in several task related to semantic similarity, e.g. text similarity, cross-level semantic similarity.

8 Conclusions and future directions

Measuring the semantic relatedness of words has long been characterized as a central component for establishing numerous cognitive processes. So, measures of similarity or relatedness have an important role in a large variety of treatments and algorithms, and are of particular interest for the development of several research fields as it is discussed in Sect. 6. The evaluation process is an important topic for the selection of context-specific measures; we are convinced that this survey provides a deep analysis to understand this topic.

[46] https://dkpro.github.io/.

[47] http://semanticsimilarity.org/.

[48] https://www.linguatools.de/disco/.

[49] http://swoogle.umbc.edu/SimService/.

[50] https://omictools.com/intego2-tool.

[51] http://xldb.di.fc.ul.pt/tools/cessm/.

[52] http://alt.qcri.org/semeval2019/.

For tasks involving lexical semantic relatedness, the difference between accuracy and complexity of the measures will be an important factor especially in large-scale tasks since calculating the semantic relationship adds an additional cost of pre-processing.

We summarized the evaluation protocol and presented an analysis based on two sides the datasets and the applications. For the datasets side, we examined the general process for the datasets construction. This process needs to be more normalized by fixing requirements and advises. We propose some recommendations in the discussion paragraph in relation with the different stages in the process. The present work consists a collection, in a large scale, of the SS/SR datasets for more then 20 languages including monolingual, multilingual and cross-lingual word similarity. We detailed, also, the metrics exploited for expressing the performance of measures for comparing the Human judgments and the automatic provided values. As for the applications,including simple derived word similarity tasks (*e.g.* short text similarity) and sophisticated applications, they made recourse to old proposals. This imposes the need for treating the replicability and reproducibility of the proposed measures, and providing the appropriate tools. It has been shown that the comparison of semantic relatedness methods is a non-trivial task. First, different methods were evaluated on different datasets, using different semantic resources and different metrics of evaluation. Second, the results of the comparison are inconclusive—no method can consistently outperform others on all datasets. The conclusion is that the choice of semantic relatedness methods should depend on a number of interrelated factors. Each category of method has certain advantages over others, but also suffers from certain limitations. These limitations are often related to the underlying basic information resources used by the methods. On the other hand, the selection of semantic relatedness methods is often limited by the availability of semantic resources in the appropriate language and the relevant fields. For this reason, it is often necessary to balance the trade-off between the potential accuracy of semantic relatedness measures and their complexity when choosing applications. In addition, we identified the remaining issues in this research field and suggested future research directions in the following.

8.1 Remaining issues

Comparative evaluation Despite the availability of benchmarking datasets for in vitro evaluation, many studies have not reported in vitro experiments, or only used the most well-known datasets. This makes it difficult to compare their methods with others and perhaps provides a partial view of the capacity of the methods. It is therefore important for future studies to make a comparative assessment in order 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 replication and comparison of experiments.

Multilingualism SS/SR measures can be applied to different languages by simply adapting methods to the background information resources of particular languages. In practice, although the performance of different methods for different languages is generally consistent, some may require specialization because of different (eg, morphological, syntactic) language characteristics. At the same time, research on the multilingual and cross-lingual word similarity is useful for its wide application.

Application-specific tasks As an alternative to the direct evaluation of SR/SS measures through a gold standard, application-specific tasks are often used to measure the impact of the proposed measures on improving the performance of a particular task. The underlying hypothesis of application-specific evaluations is that the more accurate a SR measure is, the more it improves the performance of the task at hand. The advantage of the evaluation based on specific tasks in application is that not only it shows whether the SR/SS measure is able to cause any notable improvement but also shows how well the SR/SS measure is suitable for domain specific tasks. For instance, experimentation can show that given SR method does not perform well under all conditions, it is effective for a specific task or application area.

Selection of evaluation technique In terms of evaluating the developed semantic measures, the present review shows that most authors have adopted the evaluation of correlation between automatic computed values and human similarity/relatedness judgments. One of the important factors in deciding which datasets to adopt is the inter-annotator agreement of the participants from whom the similarity values were collected. Application-specific tasks have not been widely used in the literature is that the accuracy of the semantic measure is not directly observable and is only evaluated indirectly through the performance of the higher level task. Therefore, it is possible that a good performing semantic measure is affected by the parameters inside the application framework. In order to properly use application-specific tasks for evaluation of semantic measure, a controlled experiment needs to be organized by varying only the SS/SR measures parameters.

8.2 Research directions

Better characterize semantic measures and their semantics In recent works, researchers focused on the design of semantic measures, but few contributions have focused on the semantics of these measures. In this case, the semantics to be carried by the measures is expected to be implicitly constrained by the benchmarks used to evaluate the accuracy of measures.

Designer of semantic measures are invited to provide an in-depth characterization of measures they propose. To this end, they can use, among others, the various aspects and properties of the measures detailed in this survey. Communities are encouraged to be involved in the study of semantic measures to better define what a good semantic measure is and to define exactly what makes one measure better than another. Within this goal, the study of the role of usage context seems to be of major importance. Several other properties of measures could also be taken into account and further investigated: Algorithmic complexity, degree of control on the semantics of the scores produced by the measures, the distribution of the scores produced by a measure and the reproducibility of a measure.

Develop datasets Most of the benchmarks aim at evaluating the accuracy of semantic measures according to Human judgments of similarity/relatedness. For the most part, they are composed of a reduced number of entries, e.g., pairs of words/concepts, and have been computed using a reduced pool of subjects. Initiatives for the development of benchmarks must be encouraged in order to obtain larger benchmarks in various domains of study and for different languages. Word-to-word benchmarks should take in consideration the conceptualized format for evaluating knowledge-based semantic measures. Moreover, it is important to propose benchmarks other than human judgments of similarity, i.e., benchmarks for in-vivo evaluation strategy based on the analysis of the performance of applications which rely on semantic measures. More analysis of existing benchmarks must also be performed in order to criticize their relevance and to underline their strengths and limits.

Corpus-based semantic measures and language specificities It is important to analyze semantic measures in order to ensure that accurate models of semantic similarity and semantic relatedness are available to process resources which are not expressed in English. Among others, open problems are related to: (1) the evaluation of the use of exiting models in different languages, (2) the definition of measures which are accurate for multiple languages and (3) the definition and study of language-specific processes which can be used to improve measure accuracy.

Study the algorithmic complexity of semantic measures The algorithmic complexity as a scoop in the definition of the semantic measures is near inexistent despite the fact that this aspect is essential for the efficiency of the application. However, these aspects are essential for comparing semantic measures. Indeed, in most application contexts, users will prefer to reduce measure accuracy for a significant reduction of the computational time in relation with the resources required to use a measure. To this end, designers of semantic measures are invited to provide, as possible, the algorithmic complexity of their proposals.

Examining the combination between knowledge-based measures and word embedding models Recently several works has been focused on the development of word embedding models that show good performance with the datasets. But rare works (Lastra-Díaz et al. 2019b) are examining this field. Joint approaches merge the complementary information from text and knowledge base while they compute the word embeddings and learning them from scratch. We think that this is the more appropriate time for benefiting from the enhancements made in these two families of measures.

Supporting the replicability and the reproducibility These findings confirm again the reproducibility problems in the area. In fact, the applications involving SS/SR measures in their treatment process exploited the older proposals. Thus, we invite the research community to reproduce the methods and experiments reported in the literature in order to confirm or refute the results reported herein, as well as to publish their software implementations. Therefore, several journals and known conferences open calls for papers focusing on the reproducibility aspect.

References

Abdi A, Idris N, Alguliyev RM, Aliguliyev RM (2015) Pdlk: plagiarism detection using linguistic knowledge. Expert Syst Appl 42(22):8936–8946

Agirre E, Alfonseca E, Hall K, Kravalova J, Paşca M, Soroa A (2009) A study on similarity and relatedness using distributional and wordnet-based approaches. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics, NAACL'09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 19–27

Agirre E, Diab M, Cer D, Gonzalez-Agirre A (2012) Semeval-2012 task 6: a pilot on semantic textual similarity. In: Proceedings of the first joint conference on lexical and computational semantics—volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation, SemEval'12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 385–393

Akhtar SS, Gupta A, Vajpayee A, Srivastava A, Shrivastava M (2017) Word similarity datasets for Indian languages: annotation and baseline systems. In: Proceedings of the 11th linguistic annotation workshop at ACL, pp 91–94

Akmal S, Shih LH, Batres R (2014) Ontology-based similarity for product information retrieval. Comput Ind 65(1):91–107

Alkhatlan A, Kalita J, Alhaddad A (2018) Word sense disambiguation for arabic exploiting arabic wordnet and word embedding. Proc Comput Sci 142:50–60

Almarsoomi FA, O'Shea J, Bandar Z, Crockett KA (2013) AWSS: an algorithm for measuring arabic word semantic similarity. In: IEEE international conference on systems, man, and cybernetics, Manchester, SMC 2013, United Kingdom, October 13–16, 2013, pp 504–509

Almuhareb A (2006) Attributes in lexical acquisition. Ph.D. thesis, University of Essex, England, Essex

Angelos H, Giannis V, Epimeneidis V, Euripides GMP, Evangelos M (2006) Information retrieval by semantic similarity. J Semant Web Inf Syst (IJSWIS) 3(3):55–73

Araque O, Zhu G, Garcí-Amado M, Iglesias CA (2016) Mining the opinionated web: classification and detection of aspect contexts for aspect based sentiment analysis. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW), pp 900–907

Araque O, Zhu G, Iglesias CA (2019) A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowl Based Syst 165:346–359

Artstein R (2017) Inter-annotator agreement. Handbook of linguistic annotation. Springer, Dordrecht, pp 297–313

Avraham O, Goldberg Y (2016) Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In: RepEval@ACL. Association for Computational Linguistics, pp 106–110

Baker S, Reichart R, Korhonen A (2014) An unsupervised model for instance level subcategorization acquisition. In: Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, a meeting of SIGDAT, a special interest group of the ACL, pp 278–289

Ballatore A, Bertolotto M, Wilson DC (2014) An evaluative baseline for geo-semantic relatedness and similarity. Geoinformatica 18(4):747–767

Banerjee S, Pedersen T (2003) Extended gloss overlaps as a measure of semantic relatedness. In: In Proceedings of the eighteenth international joint conference on artificial intelligence, pp 805–810

Bangalore S, Haffner P, Kanthak S (2007) Statistical machine translation through global lexical selection and sentence reconstruction. In: ACL 2007, proceedings of the 45th annual meeting of the Association for Computational Linguistics, June 23–30, 2007, Prague, Czech Republic (2007)

Bär D, Zesch T, Gurevych I (2011) A reflective view on text similarity. In: Angelova G, Bontcheva K, Mitkov R, Nicolov N (eds) RANLP. RANLP 2011 organising committee, pp 515–520 (2011)

Bär D, Biemann C, Gurevych I, Zesch T (2012a) UKP: computing semantic textual similarity by combining multiple content similarity measures. In: Proceedings of the 6th international workshop on semantic evaluation, held in conjunction with the 1st joint conference on lexical and computational semantics, pp 435–440

Bär D, Zesch T, Gurevych I (2012b) Text reuse detection using a composition of text similarity measures. In: Proceedings of the 24th international conference on computational linguistics (COLING 2012). Mumbai, India, pp 167–184. http://www.aclweb.org/anthology/C12-1011

Bär D, Zesch T, Gurevych I (2013) Dkpro similarity: an open source framework for text similarity. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics: system demonstrations. Association for Computational Linguistics, pp 121–126

Bär D, Zesch T, Gurevych I (2015) Composing measures for computing text similarity. Technical report

Baroni M, Lenci A (2011) How we BLESSed distributional semantic evaluation. In: Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics. Association for Computational Linguistics, Edinburgh, UK, pp 1–10

Baroni M, Murphy B, Barbu E, Poesio M (2010) Strudel: a corpus-based semantic model based on properties and types. Cognit Sci 34(2):222–254

Barzegar S, Sales JE, Freitas A, Handschuh S, Davis B (2015) Dinfra: a one stop shop for computing multilingual semantic relatedness. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, SIGIR'15. New York, NY, USA, pp 1027–1028

Barzegar S, Davis B, Zarrouk M, Handschuh S, Freitas A (2018) Semr-11: a multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In: Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018

Bell MJ, Schäfer M (2016) Modelling semantic transparency. Morphology 26(2):157–199

Ben Aouicha M, Hadj Taieb MA, Ibn Marai H (2016a) WSD-TIC: word sense disambiguation using taxonomic information content. In: Computational collective intelligence—8th international conference, ICCCI 2016, Halkidiki, Greece, September 28–30, 2016, proceedings, part I, pp 131–142

Ben Aouicha M, Hadj Taieb MA, Ben Hamadou A (2016b) Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Appl Intell 45(2):475–511

Ben Aouicha M, Hadj Taieb MA, Ben Hamadou A (2018a) SISR: system for integrating semantic relatedness and similarity measures. Soft Comput 22(6):1855–1879

Ben Aouicha M, Hadj Taieb M, Ibn Marai H (2018b) Wordnet and wiktionary-based approach for word sense disambiguation. Trans Comput Collective Intell 29:123–143

Bernstein A, Kaufmann E, Kiefer C, Bürki C (2005) Simpack: a generic java library for similarity measures in ontologies. Technical report

Biemann C, Riedl M (2013) Text: now in 2D! A framework for lexical expansion with contextual similarity. J Lang Model 1(1):55–95

Bird S (2006) Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, COLING-ACL'06. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 69–72

Bjerva J, Östling R (2017) Cross-lingual learning of semantic textual similarity with multilingual word representations. In: Proceedings of the 21st nordic conference on computational linguistics. Association for Computational Linguistics, pp 211–215

Blair P, Merhav Y, Barry J (2017) Automated generation of multilingual clusters for the evaluation of distributed representations. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, workshop track proceedings

Bollegala D, Matsuo Y, Ishizuka M (2007) Measuring semantic similarity between words using web search engines. In: WWW'07: proceedings of the 16th international conference on world wide web. ACM, pp 757–766

Bruni E, Tran NK, Baroni M (2014) Multimodal distributional semantics. J Artif Int Res 49(1):1–47

Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of semantic distance. Comput Linguist 32(1):13–47

Camacho-Collados J, Navigli R (2016) Find the word that does not belong: a framework for an intrinsic evaluation of word vector representations. In: Proceedings of the 1st workshop on evaluating vector-space representations for NLP. Association for Computational Linguistics, Berlin, Germany, pp 43–50

Camacho-Collados J, Pilehvar MT, Navigli R (2015) A framework for the construction of monolingual and cross-lingual word similarity datasets. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing, ACL 2015, July 26–31, 2015, Beijing, China, vol 2, pp 1–7

Camacho-Collados J, Pilehvar MT, Collier N, Navigli R (2017) Semeval-2017 task 2: multilingual and cross-lingual semantic word similarity. Vancouver, Canada

Carvalho D, Çalli C, Freitas A, Curry E (2014) Easyesa: a low-effort infrastructure for explicit semantic analysis. In: Proceedings of the 2014 international conference on posters & demonstrations track, ISWC-PD'14, vol 1272. Aachen, Germany, pp 177–180

Cer DM, Diab MT, Agirre E, Lopez-Gazpio I, Specia L (2017) Semeval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th international workshop on semantic evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3–4, 2017, pp 1–14

Chen F, Lu C, Wu H, Li M (2017) A semantic similarity measure integrating multiple conceptual relationships for web service discovery. Expert Syst Appl 67:19–31

Chen Z, Song J, Yang Y (2018) An approach to measuring semantic relatedness of geographic terminologies using a thesaurus and lexical database sources. ISPRS Int J Geo-Inf 7(3):98

Cilibrasi RL, Vitanyi PMB (2007) The google similarity distance. IEEE Trans Knowl Data Eng 19(3):370–383

Cinková S (2016) WordSim353 for czech. Springer, Cham, pp 190–197

Cohen KB, Xia J, Zweigenbaum P, Callahan T, Hargraves O, Goss F, Ide N, Névéol A, Grouin C, Hunter LE (2018) Three dimensions of reproducibility in natural language processing. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC (2018) European Language Resources Association (ELRA). Miyazaki, Japan

Curran JR (2002) Ensemble methods for automatic thesaurus extraction. In: Proceedings of conference on empirical methods in natural language processing, pp 222–229

David J, Euzenat J (2008) Comparison between ontology distances (preliminary results). In: Sheth A, Staab S, Dean M, Paolucci M, Maynard D, Finin T, Thirunarayan K (eds) The semantic web-ISWC 2008. Springer, Berlin, pp 245–260

de Saussure F (1983) Course in general linguistics. Duckworth, London ([1916] 1983). (trans. Roy Harris)

Dinu G, Pham NT, Baroni M (2013) DISSECT—DIStributional SEmantics composition toolkit. In: Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations. Association for Computational Linguistics, Sofia, Bulgaria, pp 31–36

Egozi O, Gabrilovich E, Markovitch S (2008) Concept-based feature generation and selection for information retrieval. In: Proceedings of the twenty-third AAAI conference on artificial intelligence

Ensan F, Du W (2018) Ad hoc retrieval via entity linking and semantic similarity. Knowl Inf Syst 58:551–583

Ercan G, Yildiz OT (2018) Anlamver: semantic model evaluation dataset for turkish—word similarity and relatedness. In: Proceedings of the 27th international conference on computational linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018, pp 3819–3836

Fellbaum C (ed) (1998) WordNet an electronic lexical database. The MIT Press, Cambridge

Feng Y, Bagheri E, Ensan F, Jovanovic J (2017) The state of the art in semantic relatedness: a framework for comparison. Knowl Eng Rev 32:1–30

Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, Wolfman G, Ruppin E (2002) Placing search in context: the concept revisited. ACM Trans Inf Syst 20(1):116–131

Franco-Salvador M, Rosso P, Montes-y-Gómez M (2016) A systematic study of knowledge graph analysis for cross-language plagiarism detection. Inf Process Manag 52(4):550–570

Freitas A, Barzegar S, Sales JE, Handschuh S, Davis B (2016) Semantic relatedness for all (languages): a comparative analysis of multilingual semantic relatedness using machine translation. In: Blomqvist E, Ciancarini P, Poggi F, Vitali F (eds) Knowledge engineering and knowledge management: 20th international conference, EKAW 2016, Bologna, Italy, November 19–23, 2016, Proceedings. Springer International Publishing, Cham, pp 212–222

Gabsi I, Kammoun H, Brahmi S, Amous I (2017) Mesh-based disambiguation method using an intrinsic information content measure of semantic similarity. Proc Comput Sci 112:564–573

Garla VN, Brandt C (2012) Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinform 13:261–261

Gerz D, Vulic I, Hill F, Reichart R, Korhonen A (2016) Simverb-3500: a large-scale evaluation set of verb similarity. In: Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, Austin, Texas, USA, November 1–4, 2016, pp 2173–2182

Gil JM, Montes JFA (2013) Semantic similarity measurement using historical google search patterns. Inf Syst Front 15(3):399–410

Glavas G, Nanni F, Ponzetto SP (2016) Unsupervised text segmentation using semantic relatedness graphs. In: Proceedings of the fifth joint conference on lexical and computational semantics, *SEM@ACL 2016, Berlin, Germany, 11–12 August 2016

Gliozzo A, Strapparava C (2006) Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics, ACL-44. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 553–560

Gracia J, Mena E (2008) Web-based measure of semantic relatedness. In: Proceedings of 9th international conference on web information systems engineering (WISE 2008), Auckland, New Zealand. Springer, pp 136–150

Granada R, Trojahn C, Vieira R (2014) Comparing semantic relatedness between word pairs in portuguese using wikipedia. Springer, Cham, pp 170–175

Guessoum D, Miraoui M, Tadj C (2015) Survey of semantic simialrity measures in pervasive computing. Int J Smart Sens Intell Syst 8(1):125–158

Gurevych I (2005) Using the structure of a conceptual network in computing semantic relatedness. In: Natural language processing—IJCNLP 2005, second international joint conference, Jeju Island, Korea, October 11–13, 2005, proceedings, pp 767–778

Gurevych I (2006) Computing semantic relatedness across parts of speech. Darmstadt University of Technology, Germany, Department of Computer Science, Telecooperation, technical report

Gurevych I, Strube M (2004)Semantic similarity applied to spoken dialogue summarization. In: Proceedings of the 20th international conference on computational linguistics, COLING'04

Gurevych I, Müller C, Zesch T (2007) What to be?—Electronic career guidance based on semantic relatedness. In: Proceedings of ACL. Association for Computational Linguistics, pp 1032–1039

Guzzi PH, Mina M, Guerra C, Cannataro M (2012) Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinf 13(5):569–585

Hadj Taieb MA, Ben Aouicha M, Ben Hamadou A (2013) Computing semantic relatedness using wikipedia features. Knowl Based Syst 50:260–278

Hadj Taieb MA, Ben Aouicha M, Ben Hamadou A (2014) Ontology-based approach for measuring semantic similarity. Eng Appl AI 36:238–261

Hadj Taieb MA, Ben Aouicha M, Bourouis Y (2015) FM3S: features-based measure of sentences semantic similarity. In: Hybrid artificial intelligent systems—10th international conference, HAIS 2015, Bilbao, Spain, June 22–24, 2015, proceedings, pp 515–529

Halawi G, Dror G, Gabrilovich E, Koren Y (2012) Large-scale learning of word relatedness with constraints. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, NY, USA, pp 1406–1414

Han L, Kashyap AL, Finin T, Mayfield J, Weese J (2013) Umbc_ebiquity-core: semantic textual similarity systems. In: *SEM@NAACL-HLT. Association for Computational Linguistics, pp 44–52

Harispe S, Ranwez S, Janaqi S, Montmain J (2014) The semantic measures library: assessing semantic similarity from knowledge representation analysis. In: Métais E, Roche M, Teisseire M (eds) Natural language processing and information systems. Springer, Cham, pp 254–257

Harispe S, Ranwez S, Janaqi S, Montmain J (2015) Semantic similarity from natural language and ontology analysis. Morgan & Claypool Publishers, San Rafael

Hassan S, Mihalcea R (2009) Cross-lingual semantic relatedness using encyclopedic knowledge. In: Proceedings of the 2009 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp 1192–1201. http://www.aclweb.org/anthology/D/D09/D09-1124

Hassan S, Banea C, Mihalcea R (2012) Measuring semantic relatedness using multilingual representations. In: Proceedings of the first joint conference on lexical and computational semantics—volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation, Semeval'12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 20–29

Hecht B, Carton SH, Quaderi M, Schöning J, Raubal M, Gergle D, Downey D (2012) Explanatory semantic relatedness and explicit spatialization for exploratory search. In: Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, SIGIR'12. ACM, New York, NY, USA, pp 415–424

Hill F, Reichart R, Korhonen A (2015) Simlex-999: evaluating semantic models with (genuine) similarity estimation. Comput Linguist 41(4):665–695

Hirst G, Budanitsky A (2005) Correcting real-word spelling errors by restoring lexical cohesion. Nat Lang Eng 11(1):87–111

Hliaoutakis A (2005) Semantic similarity measures in the mesh ontology and their application to information retrieval on medline. In: Technical report, Technical University of Crete (TUC), Department of Electronic and Computer Engineering

Horsmann T, Zesch T (2018) DeepTC—an extension of DKPro text classification for fostering reproducibility of deep learning experiments. In: Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018

Huang E, Socher R, Manning C, Ng A (2012) Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Jeju Island, Korea, pp 873–882

Jarmasz M, Szpakowicz S (2003) Roget's thesaurus and semantic similarity. In: Proceedings of conference on recent advances in natural language processing (RANLP 2003), pp 212–219

Jiang Y, Wang X, Zheng HT (2014) A semantic similarity measure based on information distance for ontology alignment. Inf Sci 278(Supplement C):76–87. https://doi.org/10.1016/j.ins.2014.03.021

Jin P, Wu Y (2012) SemEval-2012 task 4: evaluating chinese word similarity. In: Proceedings of the first joint conference on lexical and computational semantics, pp 374–377

Joubarne C, Inkpen D (2011) Comparison of semantic similarity for different languages using the google n-gram corpus and second-order co-occurrence measures. In: Advances in artificial intelligence—24th Canadian conference on artificial intelligence, Canadian AI 2011, St. John's, Canada, May 25–27, 2011. Proceedings, pp 216–221

Jurgens D, Stevens K (2010) The s-space package: an open source package for word space models. In: Proceedings of the ACL 2010 system demonstrations. Association for Computational Linguistics, Uppsala, Sweden, pp 30–35

Kennedy A, Hirst G (2012) Measuring semantic relatedness across languages. In: xLiTe: cross-lingual technologies workshop collocated with NIPS 2012

Kiela D, Hill F, Korhonen A, Clark S (2014) Improving multi-modal representations using image dispersion: why less is sometimes more. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Baltimore, Maryland, pp 835–841

Kipper K, Korhonen A, Ryant N, Palmer M (2007) A large-scale classification of english verbs. Lang Resour Eval 42(1):21–40

Kiritchenko S, Mohammad S (2017) Best-worst scaling more reliable than rating scales: a case study on sentiment intensity annotation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers). Association for Computational Linguistics, Vancouver, Canada, pp 465–470

Kolb P (2008) DISCO: a multilingual database of distributionally similar words. In: Storrer A, Geyken A, Siebert A, Würzner KM (eds) KONVENS 2008—Ergänzungsband: Textressourcen und lexikalisches Wissen, pp 37–44

Konopik M, Pražák O, Steinberger D (2017) Czech dataset for semantic similarity and relatedness. In: Proceedings of the international conference recent advances in natural language processing, RANLP 2017. INCOMA Ltd., Varna, Bulgaria, pp 401–406

Kozima H (1993) Computing lexical cohesion as a tool for text analysis. Technical report

Lastra-Díaz JJ, García-Serrano A, Batet M, Fernández M, Chirigati F (2017) Hesml: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Inf Syst 66:97–118

Lastra-Díaz JJ, Goikoetxea J, Hadj Taieb MA, García-Serrano A, Ben Aouicha M, Agirre E (2019a) Word similarity benchmarks of recent word embedding models and ontology-based semantic similarity measures. e-cienciaDatos, v1. http://dx.doi.org/10.21950/AQ1CVX

Lastra-Díaz JJ, Goikoetxea J, Hadj Taieb M, García-Serrano A, Ben Aouicha M, Agirre E (2019b) A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. Eng Appl Artif Intell 85:645–665

Lastra-Díaz JJ, Goikoetxea J, Hadj Taieb M, García-Serrano A, Ben Aouicha M, Agirre E (2019c) Reproducibility dataset for a large experimental survey on word embeddings and ontology-based methods for word similarity. Data Brief 26:104432

Lee JH, Kim MH, Lee YJ (1993) Information retrieval based on conceptual distance in is-a hierarchies. J Doc 49(2):188–207

Leviant I, Reichart R (2015) Judgment language matters: multilingual vector space models for judgment language aware lexical semantics. CoRR. arXiv:abs/1508.00106

Li YM, Chen CW (2009) A synthetical approach for blog recommendation: combining trust, social relation, and semantic analysis. Expert Syst Appl 36(3):6536–6547

Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X (2011) Dosim: an R package for similarity between diseases based on disease ontology. BMC Bioinf 12(1):266

Li P, Wang H, Zhu KQ, Wang Z, Wu X (2013) Computing term similarity by large probabilistic is a knowledge. In: Proceedings of the 22Nd ACM international conference on conference on information & knowledge management, CIKM'13. ACM, New York, NY, USA, pp 1401–1410

Lin F, Sandkuhl K (2008) A survey of exploiting wordnet in ontology matching. In: Bramer M (ed) IFIP AI, IFIP, vol 276. Springer, pp 341–350

Liu Q, Liu B, Zhang Y, Kim DS, Gao Z (2016) Improving opinion aspect extraction using semantic similarity and aspect associations. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA, pp 2986–2992

Liu XY, Zhou YM, Zheng RS (2007) Measuring semantic similarity in wordnet. In: 2007 international conference on machine learning and cybernetics, vol 6, pp 3431–3435

Lopez-Gazpio I, Maritxalar M, Gonzalez-Agirre A, Rigau G, Uria L, Agirre E (2017) Interpretable semantic textual similarity: finding and explaining differences between sentences. Knowl Based Syst 119:186–199

Lord P, Stevens R, Brass A, Goble C (2003) Semantic similarity measures as tools for exploring the gene ontology. In: Proceedings of pacific symposium on biocomputing, pp 601–612

Louviere JJ (1991) Best-worst scaling: a modelfor the largest difference judgments. Working paper

Luong T, Socher R, Manning C (2013) Better word representations with recursive neural networks for morphology. In: Proceedings of the seventeenth conference on computational natural language learning. Association for Computational Linguistics, Sofia, Bulgaria, pp 104–113

Madani Y, Erritali M, Bengourram J (2019) Sentiment analysis using semantic similarity and hadoop mapreduce. Knowl Inf Syst 59(2):413–436

Mandera P, Keuleers E, Brysbaert M (2017) Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. J Mem Lang 92:57–78

Marie-Francine M (2013) Similarity measures for semantic relation extraction. Université catholique de Louvain, These

McInnes BT, Pedersen T, Pakhomov SVS (2009) UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity. In: AMIA. AMIA

Meo PD, Nocera A, Terracina G, Ursino D (2011) Recommendation of similar users, resources and social networks in a social internetworking scenario. Inf Sci 181(7):1285–1305

Meyer CM, Mieskes M, Stab C, Gurevych I (2014) Dkpro agreement: an open-source java library for measuring inter-rater agreement. In: COLING (Demos). ACL, pp 105–109

Mihalcea R, Tarau P (2004) Textrank: bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing

Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on artificial intelligence—volume 1, AAAI'06. AAAI Press, pp 775–780

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. In: 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, workshop track proceedings

Mikolov T, Yih WT, Zweig G (2013b) Linguistic regularities in continuous space word representations. In: HLT-NAACL, pp 746–751

Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. Lang Cognit Process 6(1):1–28

Monz C, Dorr BJ (2005) Iterative translation disambiguation for cross-language information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 520–527

Narducci F, Palmonari M, Semeraro G (2017) Cross-lingual link discovery with TR-ESA. Inf Sci 394–395:68–87

Navigli R (2009) Word sense disambiguation: a survey. ACM Comput Surv 41(2):10:1–10:69

Nelson DL, McEvoy CL, Schreiber TA (2004) The University of South Florida free association, rhyme, and word fragment norms. Behav Res Methods Instrum Comput 36(3):402–407

Netisopakul P, Wohlgenannt G, Pulich A (2019) Word similarity datasets for thai: Construction and evaluation. CoRR. arXiv:abs/1904.04307

Nguyen KA, Schulte im Walde S, Vu NT (2018) Introducing two Vietnamese datasets for evaluating semantic models of (dis-)similarity and relatedness, pp 199–205

Nguyen HT, Duong PH, Cambria E (2019) Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowl Based Syst 182:104–842

Nie JY, Simard M, Isabelle P, Durand R (1999) Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22Nd annual international ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, pp 74–81

Och FJ, Ney H (2000) A comparison of alignment models for statistical machine translation. In: Proceedings of the 18th conference on computational linguistics—volume 2. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1086–1090

Oldakowski R, Bizer C (2005) SemMF: a framework for calculating semantic similarity of objects represented as RDF graphs. In: Poster at the 4th international semantic web conference (ISWC 2005) (2005)

Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB (2010) Semantic similarity and relatedness between clinical terms: an experimental study. Annual symposium proceedings/AMIA symposium. AMIA symposium 2010:572–576

Pakhomov SVS, Pedersen T, McInnes BT, Melton GB, Ruggieri A, Chute CG (2011) Towards a framework for developing semantic relatedness reference standards. J Biomed Inform 44(2):251–265

Panchenko A, Morozova O (2012) A study of hybrid similarity measures for semantic relation extraction. In: Proceedings of the workshop on innovative hybrid approaches to the processing of textual data, HYBRID'12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 10–18

Panchenko A, Romanov P, Morozova O, Naets H, Philippovich A, Romanov A, Fairon C (2013) Serelex: search and visualization of semantically related words. In: European conference on information retrieval. Springer, pp 837–840

Panchenko A, Ustalov D, Arefyev N, Paperno D, Konstantinova N, Loukachevitch N, Biemann C (2016) Human and machine judgements for russian semantic relatedness. In: Analysis of images, social networks and texts (AIST'2016)

Panchenko A, Ustalov D, Arefyev N, Paperno D, Konstantinova N, Loukachevitch NV, Biemann C (2017) Human and machine judgements for Russian semantic relatedness. CoRR. arXiv:abs/1708.09702

Patwardhan S, Banerjee S, Pedersen T (2003) Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th international conference on computational linguistics and intelligent text processing, Cicling'03. Springer, Berlin, pp 241–257

Patwardhan S, Pedersen T (2006) Using WordNet-based context vectors to estimate the semantic relatedness of concepts. EACL 2006 workshop making sense of sense–bringing computational linguistics and psycholinguistics together. Trento, Italy, pp 1–8

Pedersen T, Patwardhan S, Michelizzi J (2004) Wordnet::similarity: measuring the relatedness of concepts. Demonstration papers at HLT-NAACL, (2004) HLT-NAACL-demonstrations'04. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 38–41

Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG (2007) Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inf 40(3):288–299

Peng J, Li H, Liu Y, Juan L, Jiang Q, Wang Y, Chen J (2016) Intego2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. BMC Genomics 17(5):553–560

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Empirical methods in natural language processing (EMNLP), pp 1532–1543

Pesquita C, Pessoa D, Faria D, Couto F (2009) CESSM: collaborative evaluation of semantic similarity measures. JB2009: challenges in bioinformatics

Pilehvar MT, Camacho-Collados J (2019) WIC: the word-in-context dataset for evaluating context-sensitive meaning representations. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, vol 1 (long and short papers), pp 1267–1273

Pirrò G (2012) Reword: semantic relatedness in the web of data. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence, July 22–26, Toronto, Ontario, Canada, p 2012

Pirró G, Euzenat J (2010) A feature and information theoretic framework for semantic similarity and relatedness. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, PanJZ, Horrocks I, Glimm B (eds) Proceedings of the 9th international semantic web conference (ISWC2010), Lecture notes in computer science, vol 6496. Springer, pp 615–630

Ponzetto SP, 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. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 192–199

Postma M, Vossen P (2014) What implementation and translation teach us: the case of semantic similarity measures in wordnets. In: Proceedings of the seventh global wordnet conference, pp 133–141

Radinsky K, Agichtein E, Gabrilovich E, Markovitch S (2011) A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th international conference on World Wide Web, WWW'11. ACM, New York, NY, USA, pp 337–346

Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. ELRA, Valletta, Malta, pp 45–50

Resnik P, Diab M (2000) Measuring verb similarity. In: Proceedings of the twenty-second annual conference of the cognitive science society: August 13–15 (2000) Institute for Research in Cognitive Science. University of Pennsylvania, Philadelphia, PA

Resnik P, Lin J (2010) Evaluation of NLP systems. Wiley, Hoboken, pp 271–295. https://doi.org/10.1002/9781444324044.ch11

Riloff E, Schafer C, Yarowsky D (2002) Inducing information extraction systems for new languages via cross-language projection. In: Proceedings of the 19th international conference on computational linguistics—volume 1, COLING'02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1–7

Rubenstein H, Goodenough JB (1965) Contextual correlates of synonymy. Commun ACM 8(10):627–633

Ruiz-Casado M, Alfonseca E, Castells P (2005) Using context-window overlapping in synonym discovery and ontology extension. In: International conference on recent advances in natural language processing (RANLP 2005), Borovets, Bulgaria

Rus V, Lintean MC, Banjade R, Niraula NB, Stefanescu D (2013) Semilar: the semantic similarity toolkit. In: ACL (conference system demonstrations). The Association for Computer Linguistics, pp 163–168

Saad M, Langlois D, Smaïli K (2014) Cross-lingual semantic similarity measure for comparable articles. In: Przepiórkowski A, Ogrodniczuk M (eds) Adv Nat Lang Process. Springer, Cham, pp 105–115

Sahami M, Heilman TD (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. ACM, New York, NY, USA, pp 377–386

Sahlgren M (2006) The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. thesis, Stockholm University, Stockholm, Sweden

Saif A, Aziz M, Omar N (2014) Evaluating knowledge-based semantic measures on arabic. Int J Commun Antenna Propag 4(5):180–194

Sakaizawa Y, Komachi M (2017) Construction of a Japanese word similarity dataset. CoRR. arXiv:abs/1703.05916

Salem A, Ben-Abdallah H (2015) The design of valid multidimensional star schemas assisted by repair solutions. Vietnam J Comput Sci 2(3):169–179

Sales JE, Souza L, Barzegar S, Davis B, Freitas A, Handschuh S (2018) Indra: a word embedding and semantic relatedness server. In: Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018

Sánchez D, Moreno A (2008) Learning non-taxonomic relationships from web documents for domain ontology construction. Data Knowl Eng 64(3):600–623

Sánchez D, Isern D, Millan M (2011) Content annotation for the semantic web: an automatic web-based approach. Knowl Inf Syst 27(3):393–418

Santus E, Wang H, Chersoni E, Zhang Y (2018) A rank-based similarity metric for word embeddings. In: Proceedings of the 56th annual meeting of the association for computational linguistics vol 2 (short papers). Association for Computational Linguistics, Melbourne, Australia, pp 552–557

Šarić F, Glavaš G, Karan M, Šnajder J, Bašić BD (2012) Takelab: systems for measuring semantic text similarity. In: Proceedings of the first joint conference on lexical and computational semantics—volume 1: proceedings of the main conference and the shared task, and volume 2: proceedings of the sixth international workshop on semantic evaluation, SemEval'12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 441–448

Schickel-Zuber V, 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. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007)

Schuler KK (2005) Verbnet: a broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Philadelphia, PA, USA

Sen S, Li TJJ, Team W, Hecht B (2014) Wikibrain: democratizing computation on wikipedia. In: Proceedings of the international symposium on open collaboration, OpenSym'14. ACM, New York, NY, USA, pp 27:1–27:10

Silberer C, Lapata M (2014) Learning grounded meaning representations with autoencoders. In: Proceedings of the 52nd annual meeting of the association for computational linguistics vol 1 (long papers). Association for Computational Linguistics, Baltimore, Maryland, pp 721–732

Sopaoglu U, Ercan G (2016) Evaluation of semantic relatedness measures for Turkish language. In: CICLing (1), lecture notes in computer science, vol 9623. Springer, pp 600–611

Srihari RK, Zhang Z, Rao A (2000) Intelligent indexing and semantic retrieval of multimodal documents. Inf Retr 2(2):245–275

Szumlanski SR, Gomez F, Sims VK (2013) A new set of norms for semantic relatedness measures. In: ACL (2). The Association for Computer Linguistics, pp 890–895

Tan BV, Thai NP, Lam PV (2017) Construction of a word similarity dataset and evaluation of word similarity techniques for Vietnamese. In: 9th international conference on knowledge and systems engineering (KSE), pp 65–70

Torsten Z, Iryna G (2006) Automatically creating datasets for measures of semantic relatedness. Coling/ACL 2006 workshop on linguistic distances. Australia, Sydney, pp 16–24

Tóth Á (2013) How similar: word similarity judgments in english and Hungarian. Technical report

Tsatsaronis G, Varlamis I, Vazirgiannis M (2010a) Text relatedness based on a word thesaurus. J Artif Int Res 37(1):1–40

Tsatsaronis G, Giannakoulopoulos A, Varlamis I, Kanellopoulos N (2010b) Identifying free text plagiarism based on semantic similarity. In: Proceedings of the 4th international plagiarism conference. Newcastle upon Tyne, UK

Uddin MN, Duong TH, Nguyen NT, Qi XM, Jo GS (2013) Semantic similarity measures for enhancing information retrieval in folksonomies. Expert Syst Appl 40(5):1645–1653

Vulic I, Moens M (2013) Cross-lingual semantic similarity of words as the similarity of their semantic word responses. Human language technologies: conference of the north American chapter of the association of computational linguistics, proceedings, June 9–14, 2013. Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pp 106–116

Vulic I, Moens M (2014) Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In: Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, a meeting of SIGDAT, a Special Interest Group of the ACL, pp 349–362

Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinform 23(10):1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Wang X, Jia Y, Zhou B, Ding ZY, Liang Z (2011) Computing semantic relatedness using Chinese wikipedia links and taxonomy. J Chin Comput Syst 32(11):2237–2242

Wang S, Huang C, Yao Y, Chan A (2015) Mechanical turk-based experiment vs laboratory-based experiment: a case study on the comparison of semantic transparency rating data. In: Proceedings of the

29th Pacific Asia conference on language, information and computation, PACLIC 29, Shanghai, China, October 30–November 1, 2015

Wang B, Wang A, Chen F, Wang Y, Kuo CCJ (2019a) Evaluating word embedding models: methods and experimental results. APSIPA Trans Signal Inf Process. https://doi.org/10.1017/ATSIP.2019.12

Wang Y, Wang M, Fujita H (2019b) Word sense disambiguation: a comprehensive knowledge exploitation framework. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.105030

Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE (2009) Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol 7(11):e1000247

Weeds J (2003) Measures and applications of lexical distributional similarity. Ph.D. thesis, Department of Informatics, University of Sussex

Wieling M, Rawee J, van Noord G (2018) Reproducibility in computational linguistics: are we willing to share? Comput Linguist 44(4):641–649

Wu Y, Li W (2016) Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement. In: Natural language understanding and intelligent applications—5th CCF conference on natural language processing and chinese computing, NLPCC 2016, and 24th international conference on computer processing of oriental languages, ICCPOL 2016, Kunming, China, December 2–6,2016, proceedings, pp 828–839

Xie S, Liu Y (2008) Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, ICASSP 2008, March 30–April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA, pp 4985–4988

Xie F, Wu X, Hu X (2010) Keyphrase extraction based on semantic relatedness. In: Proceedings of the 9th IEEE international conference on cognitive informatics, ICCI 2010, July 7–9, 2010, Beijing, China, pp 308–312

Yang D, Powers DMW (2006) Verb similarity on the taxonomy of wordnet. In: The 3rd international WordNet conference (GWC-06), Jeju Island, Korea

Yang X, Su J (2007) Coreference resolution using semantic relatedness information from automatically discovered patterns. In: ACL. The Association for Computational Linguistics

Zesch T (2010) Study of semantic relatedness of words using collaboratively constructed semantic resources. Ph.D. thesis, Darmstadt University of Technology

Zesch T (2012) Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics (EACL 2012). Avignon, France, pp 529–538

Zesch T, Gurevych I (2010) Wisdom of crowds versus wisdom of linguists–measuring the semantic relatedness of words. Nat Lang Eng 16(1):25–59

Zhang SB, Tang QR (2016) Protein-protein interaction inference based on semantic similarity of gene ontology terms. J Theor Biol 401:30–37

Zhang Z, Gentile A, Ciravegna F (2012) Recent advances in methods of lexical semantic relatedness–a survey. Nat Lang Eng 1(1):1–69

Zhu G, Iglesias CA (2017) Sematch: semantic similarity framework for knowledge graphs. Knowl Based Syst 130:30–32

Zhu G, Iglesias CA (2018) Exploiting semantic similarity for named entity disambiguation in knowledge graphs. Expert Syst Appl 101:8–24

Ziegler CN, Simon K, 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. ACM, New York, NY, USA, pp 465–474

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.