Engineering Applications of Artificial Intelligence 85 (2019) 645–665

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art ☆

Juan J. Lastra-Díaz [a,*], Josu Goikoetxea [b], Mohamed Ali Hadj Taieb [c], Ana García-Serrano [a], Mohamed Ben Aouicha [c], Eneko Agirre [b]

[a] NLP & IR Research Group, ETSI de Informática (UNED) Universidad Nacional de Educación a Distancia, Juan del Rosal 16, 28040 Madrid, Spain
[b] IXA NLP group, Faculty of Informatics, UPV/EHU Manuel Lardizabal 1 (20018), Donostia, Basque Country, Spain
[c] Faculty of Sciences of Sfax, Tunisia

ARTICLE INFO

Keywords:
Ontology-based semantic similarity measures
Word embedding models
Information Content models
WordNet
Experimental survey
HESML

ABSTRACT

Human similarity and relatedness judgements between concepts underlie most of cognitive capabilities, such as categorisation, memory, decision-making and reasoning. For this reason, the proposal of methods for the estimation of the degree of similarity and relatedness between words and concepts has been a very active line of research in the fields of artificial intelligence, information retrieval and natural language processing among others. Main approaches proposed in the literature can be categorised in two large families as follows: (1) Ontology-based semantic similarity Measures (OM) and (2) distributional measures whose most recent and successful methods are based on Word Embedding (WE) models. However, the lack of a deep analysis of both families of methods slows down the advance of this line of research and its applications. This work introduces the largest, reproducible and detailed experimental survey of OM measures and WE models reported in the literature which is based on the evaluation of both families of methods on a same software platform, with the aim of elucidating what is the state of the problem. We show that WE models which combine distributional and ontology-based information get the best results, and in addition, we show for the first time that a simple average of two best performing WE models with other ontology-based measures or WE models is able to improve the state of the art by a large margin. In addition, we provide a very detailed reproducibility protocol together with a collection of software tools and datasets as supplementary material to allow the exact replication of our results.

1. Introduction

Measuring semantic similarity and relatedness between concepts or words is an important task in many fields of research, such as knowledge management (Ben Aouicha et al., 2016d; Georgiev and Georgiev, 2018), information retrieval (Ji et al., 2017), artificial intelligence (Liu et al., 2016), natural language processing (Hadj Taieb et al., 2015; Wu et al., 2017), biomedical domains (Ben Aouicha and Hadj Taieb, 2016), web service discovery (Chen et al., 2017), building knowledge graphs (Zhu and Iglesias, 2017), named entity disambiguation (Zhu and Iglesias, 2018), word sense disambiguation (Ben Aouicha et al., 2016e) and cross-lingual text similarity (Glavas et al., 2018) among others. Human beings consider two concepts semantically close if they share a certain meaning. According to Cruse (1986), a semantic relation is the relation connecting concepts in order to highlight the links of shared significance. The notion on semantic similarity focuses on semantically similar concepts which tend to share a number of properties. For instance, *car* and *bike* are similar because both concepts are vehicles. On the other hand, semantically related concepts may not have many properties in common but have at least one classical or non-classical relationship between them which makes them semantically close. For example, *wheel* and *car* are semantically connected through the meronymy relationship.

The aim of any semantic similarity measure is to estimate the degree of resemblance between two concepts, whilst semantic relatedness measures estimate their degree of relatedness by considering any kind of relationship linking them. For instance, the concepts *car* and *fuel* have a low degree of similarity but a high degree of relationship. Semantic similarity measures only consider 'is-a' relationships between concepts, whilst semantic relatedness measures consider a wide range

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2019.07.010.

* Corresponding author.
E-mail addresses: jlastra@invi.uned.es (J.J. Lastra-Díaz), josu.goikoetxea@ehu.eus (J. Goikoetxea), mohamedali.hadjtaieb@gmail.com (M.A Hadj Taieb), agarcia@lsi.uned.es (A. García-Serrano), mohamed.benaouicha@fss.usf.tn (M.B. Aouicha), e.agirre@ehu.eus (E. Agirre).

https://doi.org/10.1016/j.engappai.2019.07.010
Received 18 February 2019; Received in revised form 10 June 2019; Accepted 7 July 2019
Available online 1 August 2019

of relationships, including classic relations such as hypernymy, hyponymy, meronymy, antonymy, synonymy; as well as other non-classic or implicit relationships which are manifested by some form of co-occurrence of words. For this reason, the automatic estimation of the degree of semantic relatedness between concepts is considered much more difficult than the estimation of the degree of similarity. Likewise, the term "semantic distance" has been also used in the literature to refer the dissimilarity between concepts. On the other hand, Tversky (1977) points out that sometimes semantic similarity should be considered asymmetric because the similarity of a concept to any of its subsumer concepts is usually considered greater than the opposite. For instance, *pear* is like a *fruit* makes more sense that *fruit* is similar to a *pear*. However, the literature mainly deals with notions of symmetrical semantic similarity and relatedness.

Most semantic similarity measures reported in the literature have been mainly based on the use of hand-crafted ontologies as WordNet (WN) which explicitly encode 'is-a' relationships between concepts, the so-called ontology-based semantic similarity measures, whilst semantic relatedness measures have been mainly based on the *distributional similarity of terms* Weeds (2003) whose main idea is to use the co-occurrence of words as a proxy of their semantic relatedness. Distributional similarity relies on the well-known *distributional hypothesis* introduced by Harris (1954) which sets that words sharing semantic relationships tend to occur in similar contexts. However, *distributional similarity of terms* does not strictly respect the notion of semantic relatedness because it is based on a statistical analysis of texts without making use of explicit semantic relations as is the case for well-defined semantic relatedness.

Common belief in the literature on the methods for the estimation of the degree of semantic similarity and relatedness between word or concept pairs is that ontology-based semantic similarity measures outperform corpus-based semantic measures in word similarity tasks, whilst the situation is reverted in the case of word relatedness tasks as pointed-out by Hadj Taieb et al. (2014b, Table 2), Lastra-Díaz and García-Serrano (2015b, §1.1) and Ben Aouicha et al. (2016a, Tables 7 and 8). In addition, corpus-based measures provide a broader lexical coverage than the ontology-based ones without their known limitations associated to the building of ontologies, such as the demand of domain experts, the difficulties in setting universally accepted concepts and relationships, their limited lexical coverage, and the difficulties in their upgrading. On the other hand, the achievements of the recent family of WE models, whose pioneering work is introduced by Mikolov et al. (2013a), are changing this former belief by bridging the gap between both families of methods in word similarity tasks. Current WE models have significantly improved the performance of previous distributional measures (Nalisnick et al., 2016), and they are also challenging ontology-based semantic similarity measures in word similarity benchmarks as shown by Auguste et al. (2017) and Banjade et al. (2015), despite that ontology-based measures has been the predominant solution for this later task during the last three decades.

The aim of this paper is to introduce a very large and detailed experimental survey of ontology-based semantic similarity measures and WE models based on the evaluation of both families of methods in most known datasets on a same software platform, with the aim of elucidating what is the state of the problem. This survey evaluates most ontology-based semantic measures based on WordNet reported in the literature during the last three decades, as well as most recent WE models and some recent hybrid methods which combine WE models with the use of ontologies. In addition, we provide a very detailed reproducibility protocol together with a collection of software tools and datasets as supplementary material which allow that all our experiments and results to be reproduced exactly.

1.1. Motivation and research questions

Our main motivation is to evaluate and compare the two main families of methods for the estimation of semantic similarity and relatedness between words and concepts with the aim of answering the following Research Questions (RQ):

- RQ1 Has been the family of OM measures definitively outperformed by state-of-the-art WE models in the estimation of the degree of (a) similarity and (b) relatedness between words?.
- RQ2 Has been decisive the use of WordNet into recent WE models to outperform previous OM and WE methods in the word similarity task?.
- RQ3 What are the current state-of-the-art methods in the semantic word similarity and relatedness tasks?.
- RQ4 Could a linear combination of two methods significantly improve their individual performance?

A second motivation is the lack of a recent and exhaustive experimental survey comparing the performance of the two main families of methods on word similarity and relatedness estimation, as well as the lack of an updated survey comparing the most recent WE models reported in the literature.

And finally, a third motivation is the lack of a fully automatic, reproducible and extensible collection of word similarity and relatedness benchmarks including most ontology-based semantic similarity measures and the most recent state-of-the-art WE models which is based on a same software platform.

1.2. Definition of the problem and contributions

Main research problem tackled by this work is the implementation of a very large and exhaustive experimental survey on semantic word similarity between the families of ontology-based semantic similarity measures and WE models on a same software platform with the aim of answering our main research questions and setting the new state-of-the-art of the problem in a very conclusive manner. Likewise, this work tackles the problem on building a very detailed and self-contained reproducibility package which allows to reproduce all methods, experiments and results detailed herein exactly. In addition, we explore the impact of linear combinations of different semantic similarity measures as a potential method for the enhancement of their performance.

The rest of the paper is structured as follows. Section 2 introduces a comprehensive and updated categorisation of the family of ontology-based semantic similarity measures, whilst Section 3 reviews the literature on the family of WE models. Section 4 introduces our experimental setup and the results obtained in our experiments, whilst Section 5 introduces our discussion of the results and Section 6 summarises our main conclusions. Finally, Appendix A introduces the evaluation of the averaged measure pairs together with the statistical significance analysis comparing WE and ontology-based vector models, whilst Appendix B introduces a very detailed reproducibility package which allows that all our experiments and results to be reproduced exactly. Both aforementioned appendices are provided as supplementary material as detailed in Appendix C.

2. Ontology-based semantic similarity measures

This section reviews the literature on the family of ontology-based semantic similarity measures and introduces a comprehensive categorisation of most known methods. However, the lack of room prevents us of providing a detailed review of them. For this reason, for an in-depth review on these methods, we refer the reader to the book by Harispe et al. (2015b) and the more recent reviews by Hadj Taieb et al. (2014b) and Lastra-Díaz and García-Serrano (2015b,a, 2016).

- Ontology-based semantic similarity Measures (OM)
  - (I) Ontology-based semantic Topological similarity Measures (OTM)
    - (I.1) Path-based similarity measures
      Rada et al. (1989), Wu and Palmer (1994), Hirst and St-Onge (1998), Al-Mubaid and Nguyen (2009), Liu et al. (2007), Hao et al. (2011), Pekar and Staab (2002), Dong et al. (2010)
      Monotone mappings of the Rada et al. measure
      Leacock and Chodorow (1998), Li et al. (2003, strategy 3), Pedersen et al. (2007)
    - (I.2) IC-based similarity measures
      - (I.2.1) Classic IC-based measures
        Resnik (1995), Jiang and Conrath (1997), Lin (1998)
      - (I.2.2) Hybrid measures combining path-based features with an IC model
        Li et al. (2003), Zhou et al. (2008b), Meng et al. (2014), Hadj Taieb et al. (2014a), Gao et al. (2015, strategy 3), Lastra-Díaz and García-Serrano (2015b, coswJ&C), Cai et al. (2017, strategy 2),
      - (I.2.3) IC-based reformulations of previous methods
        Pirró and Seco (2008), Pirró (2009), Sánchez and Batet (2011)
      - (I.2.4) IC-based monotone mappings
        - Based on the Jiang and Conrath measure: Garla and Brandt (2012), Lastra-Díaz and García-Serrano (2015b, cosJ&C)
        - Based on the Lin measure: Pirró and Euzenat (2010), Meng et al. (2012),
      - (I.2.5) IC-based measure based on Linked Open Data (LOD)
        Meymandpour and Davis (2016)
    - (I.3) Feature-based similarity measures
      Tversky (1977), Petrakis et al. (2006), Rodríguez and Egenhofer (2003), Stojanovic et al. (2001), Batet et al. (2011), Sánchez et al. (2012), Likavec et al. (2019)
    - (I.4) Graph-based measures based on WordNet
      Stanchev (2014), Quintero et al. (2019), Cai et al. (2018)
    - (I.5) Other recent approaches
      - (I.5.1) Novel taxonomical features
        Hadj Taieb et al. (2014b)
      - (I.5.2) Aggregation or combination of other measures
        Martinez-Gil (2016), Batet and Sánchez (2016)
      - (I.5.3) Similarity between Wordnet adjectives
        Van Miltenburg (2016)
      - (I.5.4) Fuzzy logic and WordNet
        Manna and Mendis (2010)
  - (II) Gloss-based semantic similarity measures
    - Lesk (1986)
    - Based on WordNet glosses: Banerjee and Pedersen (2003), Patwardhan and Pedersen (2006)
    - IC-based measure based on WordNet and Wikipedia glosses: Jiang et al. (2017), Hadj Taieb et al. (2012)
  - (III) Ontology-based Vector Models (OVM)
    - (III.1) Based on a single ontology
      - Agirre and Soroa (2009) (UKB)
      - Goikoetxea et al. (2015) (WN-random-walks)
      - Other methods not evaluated herein: Mihalcea (2005), Sinha and Mihalcea (2007), Hassan and Mihalcea (2011), Witten and Milne (2008), Banerjee and Pedersen (2002), Budanitsky and Hirst (2006), Banko et al. (2007), Patwardhan and Pedersen (2006)
    - (III.2) Based on multiple ontologies
      Camacho-Collados et al. (2016) (NASARI) based on BabelNet

Fig. 1. Categorisation of the main ontology-based semantic similarity measures reported in the literature in the field of NLP, especially those based on WordNet. We exclude most specific GO-based semantic similarity measures proposed in the field of genomics which are in-depth analysed in a recent survey by Mazandu et al. (2016).

Fig. 1shows our categorisation of the current ontology-based semantic similarity measures into three large families as follows: (I) Ontology-based semantic Topological Measures (OTM), whose main feature is that their computing method only uses features derived from the structure (topology) of the underlying base ontology; (II) gloss-based semantic similarity measures, whose main feature is that they use the glosses accompanying the concepts within an ontology to build representations of their meaning; and finally, (III) Ontology-based Vector representation Models (OVM) whose main feature is that they build vectors to encode the meaning of the concepts within an ontology, then they use any vector-based similarity metric, such as the cosine function, to compute the degree of similarity between concepts.

2.1. Ontology-based semantic topological similarity measures

Ontology-based topological measures can be divided into five families depending on the main type of features used in their definition as shown in Fig. 1. First, path-based measures (I.1), the so-called edge-counting measures, whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of semantic distance, such as the pioneering work of Rada et al. (1989). Second, the family of IC-based similarity measures (I.2), whose core idea is the use of an Information Content (IC) model, such as the pioneering work of Resnik (1995). Third, the family of feature-based similarity measures (I.3), whose core idea is the use of set-theory operators between the feature sets of the concepts, such as Tversky (1977). Fourth, a pair of similarity measures based on similarity graphs (I.4) derived from WordNet introduced by Stanchev (2014), or the weighting of concepts as that introduced by Quintero et al. (2019). And fifth, other similarity measures (I.5) which are based on novel quantification methods of the hyponym set as that introduced by Hadj Taieb et al. (2014b), aggregation or combinations of other measures such as those methods introduced by Martinez-Gil (2016) and Batet and Sánchez (2016), another similarity measure for adjectives based on WordNet proposed by Van Miltenburg (2016), and finally other approach combining fuzzy logic and WordNet proposed by Manna and Mendis (2010). The main publicly available software libraries focusing on the implementation of ontology-based similarity measures based on WordNet are WordNet::Similarity (Pedersen et al., 2004) and WS4J (Shima, 2011), whose development is more stable, and the more recent SML (Harispe et al., 2014), WNetSS (Ben Aouicha et al., 2016b) and HESML (Lastra-Díaz et al., 2017) libraries which include most recent methods reported in the literature.

2.1.1. Path-based similarity measures

Rada et al. (1989) introduces the first ontology-based semantic distance measure between concepts which is defined as the length of the shortest path between concepts within an ontology. Similar ideas of Rada et al. are subsequently followed by other works, as those introduced by Wu and Palmer (1994), Hirst and St-Onge (1998), Al-Mubaid and Nguyen (2009), Liu et al. (2007), Hao et al. (2011), Pekar and Staab (2002), and Dong et al. (2010). Likewise, several path-based measures shown in Fig. 1 can be categorised as monotone transformations of the Rada et al. (1989) measure, such as shown by Lastra-Díaz and García-Serrano (2016) in a theoretically and empirically manner and confirmed herein (see Spearman correlation values in Table 5), among which we have the measures introduced by Leacock and Chodorow (1998), Li et al. (2003) and Pedersen et al. (2007).

2.1.2. IC-based semantic similarity measures

Resnik (1995) introduces the first semantic similarity measure based on an Information Content (IC) model with the aim of solving the problem on uniform weighting in the family of path-based measures. However, Resnik measure only considers the IC value of the lowest ancestor concept in its computation instead of the overall information on the path linking both input concepts. This later drawback encourages the proposal of an IC-based semantic distance by Jiang and Conrath (J&C) (1997) and an IC-based similarity measure by Lin (1998).

Information content-based approach is inspired from the Shannon (1948) theory and its core idea is that the similarity between concepts in a given ontology can be modelled by a function of the information content that both concepts have in common. The main hypothesis behind all the IC-based similarity measures is that the more abstract concepts should have a lower information content than the more specific ones. Any IC-based similarity measure is defined by the combination of one computing method and one specific intrinsic or corpus-based IC model as those mentioned below. Given a single-root taxonomy of concepts $(C, \leq_C, \Gamma)$, where Γ is the root node, the IC value of any concept $c_i \in C$ is defined by $IC(c_i) = -\log_2(p_i)$, being p_i the occurrence probability of c_i. Likewise, given a taxonomy of concepts $(C, \leq_C, \rho)$, an IC model is a positive real-valued function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$ which satisfies the next properties: (1) $IC\left(c_i\right) = -\log_2\left(p_i\right)$, (2) $IC\left(\Gamma\right) = 0$, and (3) $\forall c_i \leq_C c_j \Rightarrow IC\left(c_i\right) \geq IC\left(c_j\right)$ (monotonicity).

We have divided the family of IC-based similarity measures into five subfamilies as shown in Fig. 1. First group (I.2.1) is made up by the aforementioned three classic IC-based measures introduced by Resnik (1995), Jiang and Conrath (1997), and Lin (1998). A second group (I.2.2), called hybrid IC-based measures in Fig. 1, is made up by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work by Li et al. (2003) and other subsequent works by Zhou et al. (2008b), Meng et al. (2014), Hadj Taieb et al. (2014a), Gao et al. (2015, strategy 3), Cai et al. (2017, strategy 2) and Lastra-Díaz and García-Serrano (2015b, coswJ&C). A third group of IC-based measures is based on IC-based reformulations of previous approaches (I.2.3), such as the IC-based reformulations of the Tversky measure by Pirró (2009), and the IC-based reformulation of most edge-counting methods introduced by Sánchez and Batet (2011). A fourth group of IC-based measures (I.2.4), called monotone mappings in Fig. 1, is characterised by being monotone transformations of any classic IC-based similarity measure, such as the exponential-like scaling of the Lin measure introduced by Meng and Gu (2012), the reciprocal similarity measure of the J&C distance introduced by Garla and Brandt (2012), another exponential-like normalisation of the J&C distance introduced by Lastra-Díaz and García-Serrano (2015b, cosJ&C), and finally the monotone transformation of the Lin measure called FaITH (Pirró and Euzenat, 2010). Finally, a fifth group of measures is based on the definition of IC-based measures and IC models on Linked Open Data (LOD) resources (I.2.5), such as the work by Meymandpour and Davis (2016). Like the case of the aforementioned monotone transformations of the Rada et al. measure, the monotonicity relationship between the classic IC-based measures and their corresponding monotone mappings is also shown in a previous work (Lastra-Díaz and García-Serrano, 2016) and confirmed herein (see Spearman correlation values in Table 5).

Information content models based on WordNet. The first known IC model is based on corpus statistics and was introduced by Resnik (1995, 1999). The main drawback of the corpus-based IC models is the difficulty in getting a well-balanced and disambiguated corpus for the estimation of the concept probabilities. To bridge this gap, Seco et al. (2004) introduce the first intrinsic IC model in the literature, whose core hypothesis is that the IC models can be directly computed from intrinsic taxonomical features. Thus, the development of new intrinsic IC-based similarity measures is divided into two sub-problems: the proposal of new intrinsic IC models, and the proposal for new IC-based similarity measures. Among the main intrinsic and corpus-based IC models proposed in the literature, we find the proposals by Zhou et al. (2008a), Sebti and Barfroush (2008), Blanchard et al. (2008), Sánchez et al. (2011), Sánchez and Batet (2012), Meng et al. (2012), Yuan et al. (2013), Hadj Taieb et al. (2014a), Lastra-Díaz and García-Serrano (2015a, 2016), Adhikari et al. (2015), Ben Aouicha and Hadj Taieb (2016), Ben Aouicha et al. (2016c), Harispe et al. (2015a), Cai et al. (2017), Zhang et al. (2018) and Batet and Sánchez (2019). Other

researchers have also proposed other IC models based on WordNet focused on the estimation of the degree of relatedness between concepts, such as those introduced by Seddiqui and Aono (2010) and Pirró and Euzenat (2010). Finally, in another recent work, Jiang et al. (2017) introduce a new intrinsic IC model based on the Wikipedia category structure. Most known IC models are implemented by HESML (Lastra-Díaz et al., 2017).

2.1.3. Feature-based measures

Main hypothesis of this family of measures is the classic notion of concepts in formal logic which states that any concept could be defined by a collection of features or attributes verifiable by a logic predicate. Thus, the degree of similarity between concepts is defined by a ratio of common and distinct features. Tversky (1977) introduces the first feature-based semantic similarity measure, which is defined by a weighted variant for the complement of the symmetric difference between the feature set of two concepts. However, one drawback of classic Tversky measure is the difficulty of getting taxonomies of concepts which include an explicit definition of the feature set defining each concept. For this reason, most feature-based measures exploit multiple sources of information with the aim of inferring these missing feature sets. For instance, Sánchez et al. (2012) introduces a feature-based dissimilarity measure which is based on the use of the common ancestors between concepts as a measure of their degree of similarity. The core idea behind the Sánchez et al. measure is that the ratio of overlap between common ancestors could be interpreted as an estimation of the ratio of common features between concepts, according to the Tversky model.

Feature-based measures attempt to exploit and to aggregate between the hierarchical and content properties of the ontology to obtain the similarity or relatedness values. These type of measures are exploited both for semantic similarity and semantic relatedness. The hierarchical aspect includes the semantic relations, the typological parameters of a concept, the neighbourhood, etc. As for the content aspect, it is based on the assumption that each concept is described by a set of well-selected words indicating its meaning, such as their "glosses" in WordNet. When two concepts have more common characteristics and less non-common characteristics, they are more similar. Main features used are the available information in WordNet such as the set of synonyms, definitions (i.e., glosses) and different kinds of taxonomical or semantic relationships. Rodríguez and Egenhofer (2003) propose a similarity measure which exploits the weighted sum of similarities between synsets, features (e.g., meronyms, attributes, etc.) and neighbour concepts of evaluated concepts. Petrakis et al. (2006) propose the *X-Similarity* measure which is based on the overlapping between synsets and the concept's glosses extracted from WordNet (i.e., words extracted by parsing term definitions). Finally, other subsequent works introduce novel measures based on different feature extraction methods, such as those introduced by Stojanovic et al. (2001), Batet et al. (2011), Sánchez et al. (2012) and Likavec et al. (2019).

2.1.4. Graph-based measures

Among the most recent approaches, we have two different measures based on the definition of asymmetric similarity weighted graphs derived from WordNet, such as that introduced by Stanchev (2014) and Quintero et al. (2019), and another measure based on a symmetrical weighted graph based on WordNet introduced by Cai et al. (2018). In addition to the taxonomical structure from WordNet, the similarity graph proposed by Stanchev uses the definition and examples of use of the WordNet concepts as evidence on the relationships between concepts. The similarity graph is defined by a collection of oriented edges with asymmetric weights, in which the weights between parent and child concepts encode the probability that a user interested in the source node of an edge is also interested in the concept associated to the destination node. On the other hand, Quintero et al. (2019) introduce a semantic distance which is defined as the length of the shortest path between concepts in an asymmetric weighted graph whose weights are automatically refined through a relaxation process. Finally, Cai et al. (2018) introduce a semantic distance defined as the length of the shortest weighted path between two concepts on WordNet and a non-linear similarity function whose edge weights are defined by a ratio of the descendant number of hyponyms and hypernyms between child and parent concepts.

2.1.5. Other recent approaches

Fig. 1 shows other recent approaches as follows. First, a pair of measures based on novel taxonomical features (I.5.1), such as those proposed by Hadj Taieb et al. (2014b) which are based on a novel weighting of the hyponym set of a concept in WordNet. Second, two similarity measures based on the aggregation or combination of other measures (I.5.2), such as an aggregated similarity measure based on a combination of multiple ontology-based similarity measures proposed by Martinez-Gil (2016), and a semantic relatedness measure introduced by Batet and Sánchez (2016) which is based on the combination of highly-accurate ontology-based semantic similarity measures with a resemblance measure derived from corpus statistics. Third, a method to compute the semantic similarity between adjectives (I.5.3) proposed by Van Miltenburg (2016) which is based on the use of the similarity between their sets of derivational source names in WordNet. And finally (I.5.4), a novel semantic similarity measure based on fuzzy logic and WordNet proposed by Manna and Mendis (2010).

2.2. Gloss-based models

Lesk (1986) introduces the first gloss-based method to estimate the degree of similarity between words whose main hypothesis is that related word senses are defined using the same words. So, the semantic relatedness is quantified as the gloss overlaps. Subsequently, Banerjee and Pedersen (2003) propose a measure derived from Lesk measure, called Extended Gloss Overlap, which is based on the number of shared words, whilst Patwardhan and Pedersen (2006) propose the representation of concepts by Gloss Vectors derived from term glosses extracted from WordNet, and Ben Aouicha and Hadj Taieb (2015) propose a gloss-based semantic similarity measure based on a weighting mechanism applied on the nouns composing the glosses.

2.3. Ontology-based vector representation models

This section describes the second family of ontology-based methods for measuring semantic similarity which represent words and concepts as vectors which are derived from the graph-structure of the ontology. The family of ontology-based vector representation models (OVMs) can be divided into two categories shown in Fig. 1 as follows: (III.1) OVM models based on a single ontology, and (III.2) those based on multiple ontologies.

OVM models compute the meaning of words and concepts within the semantic structures of Knowledge Bases (KBs) based on graph-based methods which exploit the whole relational information from KBs as if they were graphs. Aforementioned graph-based methods for representing words are well-known in the NLP community (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009; Hassan and Mihalcea, 2011; Witten and Milne, 2008; Budanitsky and Hirst, 2006), showing to be useful in several NLP tasks such as Word Sense Disambiguation (WSD) (Budanitsky and Hirst, 2006; Agirre and Soroa, 2009), semantic similarity (Budanitsky and Hirst, 2006; Agirre et al., 2009) and Information Extraction (Banko et al., 2007). These methods model senses and words as represented in KBs, taking senses as nodes and relations between senses as edges, thus exploiting the relational information from KB structures, without any supervision nor corpus evidence.

Fig. 1 lists most OVM models reported in the literature. However, we focus our review here on the three OVM models evaluated in our

experiments because of space limitations. We selected the UKB (Agirre and Soroa, 2009), WN-RandomWalks (Goikoetxea et al., 2015) and NASARI (Camacho-Collados et al., 2016) methods because they provide publicly available word vector representations. UKB (Agirre and Soroa, 2009) and WN-RandomWalks (Goikoetxea et al., 2015) build their vector representations using WordNet as single ontology, whilst NASARI (Camacho-Collados et al., 2016) uses BabelNet, which combines information from several ontologies. Both UKB and NASARI construct high-dimensional spaces, one dimension per concept in the knowledge-base, whilst WN-RandomWalks uses an embedding method which produces low dimensional vectors.

2.3.1. OVM Models based on single ontologies

UKB Personalised PageRank vectors. UKB (Agirre and Soroa, 2009) is an OVM model based on the known PageRank (Brin and Page, 1998) algorithm, which ranks the vertices in a graph based on their structural relevance. PageRank can be understood as a random walk process which eventually assigns a rank to every node in the graph whose value represents the probability of an infinite random walk visiting the node. In each node, the random walker follows an edge with probability c, or, with probability $1 - c$ halts the random walk and jumps at random uniformly to any other node. The process is repeated for each sense in the graph, producing Personalised PageRank vectors (Agirre and Soroa, 2009; Agirre et al., 2010), PPV for short. Each dimension of a PPV corresponds to a node, and it could be understood as the probability of a random walk to end in that node. In order to obtain word representations, they do a linear combination of the corresponding senses, weighted according to the probabilities of each sense. In order to evaluate UKB herein, PPV vectors for every word in the dictionary of WordNet 3.0 with glosses were created by using the publicly available UKB software.[1]

Graph-based methods such as UKB (Agirre and Soroa, 2009) typically represent KB words with high-dimensional vectors. More recently, several authors (Perozzi et al., 2014; Tang et al., 2015; Goikoetxea et al., 2015; Grover and Leskovec, 2016) proposed to use low-dimensional vector-spaces, also called embeddings, that compress the structural information in graphs into a few hundred of dimensions. These approaches make use of several strategies when searching vicinity in graphs, such as random walks over the nodes (Perozzi et al., 2014; Goikoetxea et al., 2015), sampling nodes and optimising the likelihood of shallow-depth neighbours (Tang et al., 2015), or combining the latter with structural equivalences of nodes (Grover and Leskovec, 2016). After modelling neighbourhood, the mentioned methods encode the nodes in a low-dimensional space using methods like Skip-gram (Mikolov et al., 2013a). From these aforementioned methods, Perozzi et al. (2014) has not been applied to ontologies, and Grover and Leskovec (2016) and Tang et al. (2015) do not have publicly available embeddings. Thus, we discarded these latter methods from our experiments.

Random walk WordNet embeddings. Goikoetxea et al. (2015) propose an OVM model based on a two-step method as follows. First they perform a random walk over the WordNet graph using a Monte Carlo method (Avrachenkov et al., 2007). At each step of the walk, the method prints one of the lexicalisations of the current concept at random with the aim of creating a synthetic corpus reflecting the structure of the KB. Secondly, they feed the resulting synthetic corpora into the Skip-gram model (Mikolov et al., 2013a), thus processing it as if it was a text corpus and producing a low-dimensional vector for each word.

[1] http://ixa2.si.ehu.es/ukb/.

2.3.2. OVM models based on multiple ontologies

NASARI vectors. Unlike most KB-based techniques which exploit either WordNet or Wikipedia, NASARI (Camacho-Collados et al., 2016) vectors are based on BabelNet, a multilingual network (Navigli and Ponzetto, 2012) which merges knowledge from WordNet and Wikipedia among other KBs. NASARI represents every BabelNet concept by two vectors based on words and WordNet synsets respectively which are computed by using the weighting schema called lexical specificity on contextual information collected from Wikipedia articles and WordNet synsets linked to the concept. We use the unified NASARI representation in our experiments, which represents every BabelNet concept by a set of language-independent concepts.

3. Word embeddings

This section reviews the literature on the family of WE models which is divided into two categories as shown in Fig. 2: (1) text-based models, and (2) the recent hybrid embedding models which combine text-based models with the use of ontologies.

3.1. Text-based word embedding models

Since the distributional hypothesis was proposed by Harris (1954), large unlabelled text corpora has been often used to build word representations. In recent years, low dimensional representations known as word embeddings have derived low-dimensional representations by minimising typically loss function using Stochastic Gradient Descent (SGD) algorithm, often in the context of a neural network (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011; Turian et al., 2010; Mikolov et al., 2013a,b; Pennington et al., 2014). These so-called word embeddings have yielded state-of-the-art results in a wide variety of NLP tasks such as word similarity, analogy, PoS tagging or name-entity disambiguation (Collobert et al., 2011; Mikolov et al., 2013a; Socher et al., 2011). One limitation of previous approaches is that they work at the word-level, thus they need further processing to model morphologically rich languages or out-of-vocabulary words. Recent models (Santos and Zadrozny, 2014; Wieting et al., 2016; Kim et al., 2016; Bojanowski et al., 2016) based on character-level representations overcome this aforementioned limitation by using sub-word information.

Among the word-level text methods, we have excluded the older ones (Bengio et al., 2003; Collobert and Weston, 2008; Turian et al., 2010) because the recent works show that these methods are outperformed by newer ones (Mikolov et al., 2013a; Pennington et al., 2014). In addition, character-level methods (Santos and Zadrozny, 2014; Kim et al., 2016; Wieting et al., 2016) do not provide pre-trained embedding models with the only exception of the method introduced by Bojanowski et al. (2016). Thus, we focus our review on next text-based models evaluated herein.

word2vec. Mikolov et al. (2013a) introduce *word2vec* library which implements two word embedding models, namely Skip-gram and CBOW. The objective of Skip-gram is to predict each context word within a window around a target word by using only its representations in the model being learned, whilst CBOW model implements the opposite approach. Thus, CBOW aims to predict the current target word on the basis of the centroid of the representations for its context words. Loss function of CBOW integrates a negative sampling which has shown to being very efficient (Mikolov et al., 2013a; Goldberg and Levy, 2014) and whose aim is to reward the estimate of the probability of observed word–context pairs as well as penalising the estimate of the probability of random word–context pairs.

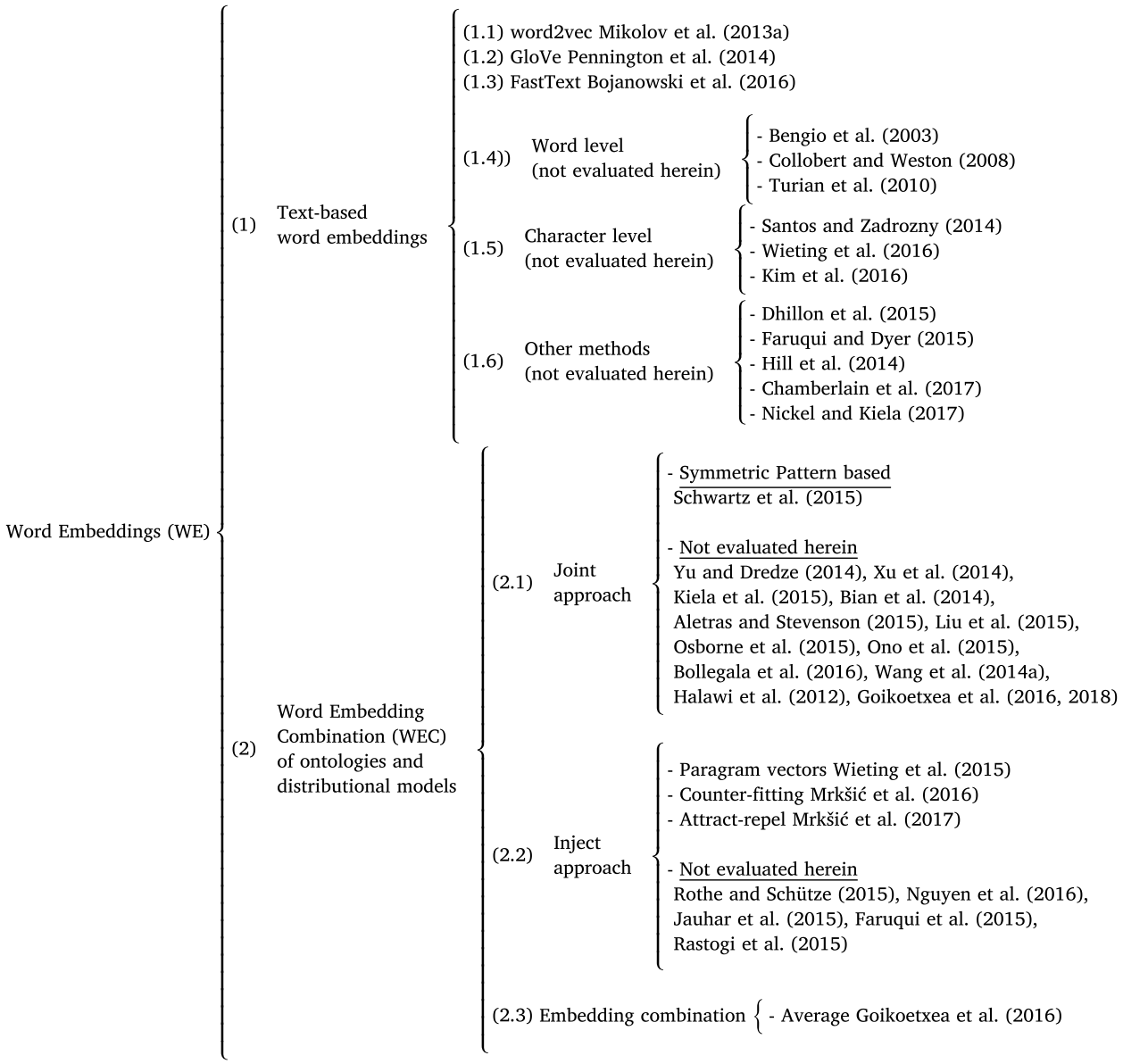

Fig. 2. Categorisation of the main Word Embedding (WE) models reported in the literature.

GloVe. Pennington et al. (2014) propose a log-bilinear model called *GloVe* which is based on a weighted least-squares objective function combining two approaches for the learning of word vectors as follows: (1) global matrix factorisation and (2) local context window methods. Pennington et al. argue that local window methods such as *word2vec* (Mikolov et al., 2013a) fail to capture global statistical information in the corpus. Thus, *GloVe* loss function minimises the least-square distance between the local window information and the global information.

fastText. Methods such as `word2vec` (Mikolov et al., 2013a,b) and `GloVe` (Pennington et al., 2014) ignore the morphology of words, and thus, they assign a vector to each word in the vocabulary. The latter is a limitation when it comes to morphologically rich languages as these kind of languages have larger vocabularies which include more word forms with lower frequency of occurrence, yielding lower quality representations. In addition, the larger vocabulary requirements produce more out-of-vocabulary words. In order to bridge this gap, Bojanowski et al. (2016) introduce *fastText* model which is based on Skip-gram but it represents each word as a bag of character n-grams. FastText learns embeddings for full words and character n-grams by representing each word as the sum of its corresponding n-gram embeddings.

3.2. Combining ontology-based and text-based representations

Several authors have exploited the complementariness of the semantic information in text-based distributional representations and relational information extracted from KBs (Halawi et al., 2012; Wang et al., 2014a; Goikoetxea et al., 2015; Faruqui et al., 2015; Goikoetxea et al., 2016; Mrkšić et al., 2016, 2017) such as WordNet (Miller, 1995), Freebase (Bollacker et al., 2008), PPDB (Ganitkevitch et al., 2013) or BabelNet (Navigli and Ponzetto, 2012). There are three main approaches as follows: (1) the joint approach, in which the loss function of the text-based model is extended to include relational constraints (Halawi et al., 2012; Wang et al., 2014a; Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Kiela et al., 2015; Aletras and Stevenson, 2015; Liu et al., 2015; Osborne et al., 2015; Ono et al., 2015; Bollegala et al., 2016; Goikoetxea et al., 2018); (2) the inject approach, which takes pre-trained text-based embeddings and transforms them to agree with the relational information in the KB (Rothe and Schütze, 2015; Faruqui

et al., 2015; Rastogi et al., 2015; Recski et al., 2016; Nguyen et al., 2016; Jauhar et al., 2015; Mrkšić et al., 2016, 2017); and finally, (3) the combination of independently learned word embeddings (Goikoetxea et al., 2016).

3.2.1. Joint approach

Joint approaches merge the complementary information from text and KB while they compute the word embeddings and learning them from scratch. These methods generally include the KB constraints in one of the WE models, or some other in-house variants (Halawi et al., 2012; Bollegala et al., 2016; Yu and Dredze, 2014; Xu et al., 2014; Bian et al., 2014; Liu et al., 2015; Ono et al., 2015; Wang et al., 2014a,b). However, other joint approaches follow different strategies. For instance, Schwartz et al. (2015) build a full co-occurrence matrix and Osborne et al. (2015) use a spectral learning algorithm, whilst Goikoetxea et al. (2016, 2018) propose an approach which combines synthetic corpora produced by WordNet random walks (Goikoetxea et al., 2015) with textual corpora.

Regarding the KB constraints, they are usually included in the objective function as L2 regularisation terms (Halawi et al., 2012; Bollegala et al., 2016) or explicit constraints encoding both entities and relations (Wang et al., 2014a; Xu et al., 2014), but also as linear combinations (Yu and Dredze, 2014), ranking inequalities (Liu et al., 2015), Canonical Correlation Analysis (Osborne et al., 2015) or multi-tasking (Bian et al., 2014). Usually, similarity relations such as synonymy (Halawi et al., 2012; Yu and Dredze, 2014) and antonymy (Ono et al., 2015; Schwartz et al., 2015) are used, but other authors (Kiela et al., 2015; Xu et al., 2014) also include relatedness ones.

Most of previous aforementioned methods have been discarded from our experiments, either because their results are outperformed by state-of-the art methods (Bian et al., 2014; Osborne et al., 2015) or because they have no publicly available pre-trained embeddings (Yu and Dredze, 2014; Liu et al., 2015; Ono et al., 2015; Wang et al., 2014a,b; Halawi et al., 2012; Goikoetxea et al., 2016), or both reasons (Xu et al., 2014; Bollegala et al., 2016). We also excluded (Goikoetxea et al., 2018) from our experiments because it has been only evaluated on cross-lingual similarity. Next paragraph reviews the only joint approach evaluated herein.

Symmetric pattern based embeddings. Methods such as word2vec (Mikolov et al., 2013a,b) consider shallow linguistic information based on word–context co-occurrences within a window. Schwartz et al. (2015) argue that these kind of approaches are suitable for relatedness, but not for similarity. Thus, they extract co-occurrences of words which occur in so-called Symmetric Patterns, e.g. “either X or Y” (Hearst, 1992; Davidov and Rappoport, 2006), as the latter text patterns tend to show semantically similar words. Likewise, Schwartz et al. (2015) use an unsupervised method to build an antonym-sensitive co-occurrence matrix which is based on the selection of word pairs satisfying antonym relations in a KB with the aim of dealing with antonym words.

3.2.2. Inject approach

In the inject approach, pre-trained word embeddings are enriched according to diverse methods. Most of inject-based methods enrich word embeddings constructed with the aforementioned methods above (Faruqui et al., 2015; Rothe and Schütze, 2015; Nguyen et al., 2016; Jauhar et al., 2015; Wieting et al., 2015; Mrkšić et al., 2017). However, Rastogi et al. (2015) construct an explicit co-occurrence matrix and Mrkšić et al. (2016) use previously enriched word embeddings. KB constraints are injected following various strategies such as similarity (Nguyen et al., 2016; Mrkšić et al., 2016, 2017; Wieting et al., 2015), weighted combinations (Faruqui et al., 2015; Jauhar et al., 2015) or Generalised Canonical Correlation Analysis (Rastogi et al., 2015). Most methods build word-based vector representations, but Rothe and Schütze (2015) extend word embeddings with embeddings of senses and synsets. As regard semantic relations, several methods only use synonymy (Faruqui et al., 2015; Rothe and Schütze, 2015), whilst other methods consider both synonymy and antonymy (Mrkšić et al., 2016, 2017) or paraphrase relations (Wieting et al., 2015). Again, the integration of these aforementioned semantic relations seek to enforce similarity rather than relatedness.

We have discarded several methods in this family from the evaluation introduced herein. For instance, the methods introduced by Faruqui et al. (2015) and Rothe and Schütze (2015) have been discarded because their lack of competitiveness, whilst methods introduced by Nguyen et al. (2016), Jauhar et al. (2015) and Rastogi et al. (2015) are discarded because their lack of publicly available pre-trained embeddings. For these reasons, we focus the rest of our review on the methods evaluated herein which are introduced by Wieting et al. (2015) and Mrkšić et al. (2016, 2017).

Paragram. Wieting et al. (2015) propose a word embedding model called Paragram whose learning method employs pairs of paraphrase phrases in PPDB (Ganitkevitch et al., 2013) database. More exactly, Paragram encodes phrases into a vector space by forcing the cosine similarity in the space to match the scores of pairs of paraphrase phrases. Wieting et al. model the composition of phrases by using constituent parse trees with using a RNN similar to those proposed by Socher et al. (2014). Paragram also includes the training of word vectors with no composition terms by using words pairs extracted from PPDB database. Finally, Wieting et al. optimise the hyperparameters of their model on a similarity dataset. We evaluate herein both versions of the publicly available Paragram embeddings (see Table 2), called Paragram-ws and Paragram-sl respectively. Paragram-ws is optimised with the similarity and relatedness partitions of WordSim353 (Finkelstein et al., 2002) dataset by rewarding vectors with high similarity and relatively low relatedness, whilst Paragram-sl is tuned with Simlex999 (Hill et al., 2014) dataset by rewarding exclusively similarity relations between word embeddings.

Counter-fitting. Similar to the Symmetric Pattern technique (Schwartz et al., 2015), this method tries to enforce similarity instead of relatedness (Mrkšić et al., 2016), using both antonymy and synonymy constraints from PPDB database and WordNet. Counter-fitting loss function is defined as the weighted sum of the three following terms: (1) a first term which ‘pushes’ away vectors of antonyms; (2) a second term which ‘pulls’ closer synonyms; and (3) a third term which forces the updated space to preserve the relationships between words in the original vector space (pre-trained embedding).

Attract–repel. Mrkšić et al. (2017) introduce the Attract–repel model which can be viewed as the cross-lingual extension of Counter-fitting. It also injects synonymy and antonymy constraints and updates pre-trained embeddings, but unlike Counter-fitting, semantic relations are drawn from BabelNet and mini-batches include negative samples in the attract and repel terms. In addition, Attract–repel uses a more straightforward L2 regularisation term to preserve word relations in the original pre-trained embeddings.

3.2.3. Embedding combination

Goikoetxea et al. (2016) propose to combine independently learned KB-based and text-based representations by using WordNet and Wikipedia graphs, although they mostly focus on the former KB. Goikoetxea et al. propose several combined representations as follows: (1) first representation combines separate distributed text-based and KB-based representations via concatenation, centroid, or building a complex number; (2) second representations exploits linear correlations, either applying Principal Component Analysis (PCA) to the concatenated representations or applying Canonical Correlation Analysis (CCA) to project both spaces into a shared space; and (3) third representation combines the scores returned by separate text and KB spaces, either as the arithmetic average or the average of the ranks. Authors report better values than retrofitting (Faruqui et al., 2014), with PCA yielding the best results, closely followed by the concatenation and average. Due to the disequilibrium in dimensionality of the word vectors in this study, we discarded all methods but the average.

4. Evaluation

The goals of the experiments described in this section are as follows: (1) a unified, reproducible and broader experimental study onto the state of the art in the families of ontology-based semantic similarity measures and WE models than previous works reported in the literature; (2) a comparison of the performance between both families of methods with the aim of answering our main research questions detailed in Section 1.1; (3) a detailed statistical significance analysis of the results; (4) the replication of previously reported methods and results; (5) a new confirmation of the achievements in both aforementioned families of semantic measures, and finally (6) the evaluation of all methods in combination with the best performing methods in noun similarity and relatedness datasets based on the average of the similarity values returned by each one.

4.1. Experimental setup

Table 1 shows the collection of ontology-based Topological similarity measures (OTM) based on WordNet which are evaluated in our experiments, whilst Table 2 shows the pre-trained word embedding (WE) and ontology-based vector (OVM) models. Table 3 details all datasets used to evaluate the methods considered herein. The selection of state-of-the-art OTM measures is based on the results obtained in four previous large experimental surveys introduced by Lastra-Díaz and García-Serrano (2015b,a, 2016) and Hadj Taieb et al. (2014b), whilst the selection of state-of-the-art WE and OVM models is based on the best performing models in the SimLex-999 Hill et al. (2015) dataset as reported in the project homepage[2]. IC-based similarity measures are evaluated in combination with their best performing IC models (Lastra-Díaz and García-Serrano, 2016, table 12), with the only exception of the Gao et al. (2015) measure which is evaluated herein with its best performing intrinsic IC model (Lastra-Díaz and García-Serrano, 2016, CondProbRefHypo). Likewise, Wu and Palmer (1994) measure evaluated herein corresponds to a commonly used approximation on tree-like taxonomies detailed by Deza and Deza (2009, p.375), instead of the original path-based measure detailed in Wu and Palmer (1994), which we call Wu&Palmer$_{fast}$.

Fig. 3 shows a concept map detailing our experimental setup to run automatically all our experiments. The evaluation of all ontology-based semantic similarity measures and WE models is based on a common software implementation provided by the release V1R4 (Lastra-Díaz and García Serrano, 2018) of the HESML library (Lastra-Díaz et al., 2017), and the noun database of WordNet 3.0 (Miller, 1995). HESML V1R4 introduces three new Java classes called EMBWordEmbeddingModel, UKBppvWordEmbeddingModel and NasariWordEmbeddingModel respectively, which implement the evaluation of the (*.emb), (*.ppv) UKB (Agirre and Soroa, 2009) and Nasari (Camacho-Collados et al., 2016) word vector file formats, with the aim of providing a common software platform for the evaluation of both aforementioned families of methods.

Vocabularies of all pre-trained embedding models shown in Table 2 are provided in lowercase. For this reason, all word pairs of the datasets detailed in Table 3 are normalised to lowercase. As consequence of the former decision, Agirre201 (Agirre et al., 2009) dataset used in our experiments differs in a few word pairs from the Agirre201 dataset evaluated by Lastra-Díaz and García-Serrano (2015a, 2016). Thus, results reported herein in the Agirre201 dataset could slightly differ in the Pearson and Spearman correlation values from those reported for the same similarity measures in the two aforementioned works. Likewise, results reported herein for the evaluation of the Leacock and Chodorow (1998) measure differ from those reported in Lastra-Díaz and García-Serrano (2016) because HESML V1R4 (Lastra-Díaz and García Serrano, 2018) fixes a software implementation error of this aforementioned measure in previous HESML versions. Finally, OTM measures are only evaluated on datasets based on noun sets contained by the noun database of WordNet 3.0, because the adjective and verb databases of WordNet 3.0 are not well-defined taxonomies. For this reason, OTM measures are only evaluated in 9 of the 19 datasets detailed in Table 3.

[2] https://www.cl.cam.ac.uk/~fh295/simlex.html.

4.2. Reproducing our benchmarks

All our experiments were generated by running the *HESMLclient* program distributed with HESML V1R4 (Lastra-Díaz and García Serrano, 2018) with an automatically reproducible benchmark file which produces a raw output file in (*.csv) file format for each dataset in Table 3 as detailed in Appendix B and shown in Fig. 3. Raw output files contain the raw similarity values returned by each semantic measure for each word pair. All tables of results reported herein are automatically generated by running the 'embeddings_vs_ontomeasures_final_tables.R' script file on the collection of raw similarity output files in the R statistical package. Finally, all our experiments and results can be exactly reproduced by following the instructions detailed in Appendix B which are based on our companion reproducibility dataset (Lastra-Díaz and García Serrano, 2018), HESML V1R4 (Lastra-Díaz and García Serrano, 2018) and Reprozip (Chirigati et al., 2016).

4.3. Evaluation metrics

As evaluation metrics, we use the Pearson correlation factor, denoted by r in Eq. (1), the Spearman rank correlation factor, denoted by ρ in Eq. (2), and the harmonic score denoted by h as defined in Eq. (3).

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{1}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \qquad di = \left(x_i - y_i\right) \tag{2}$$

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

The Pearson correlation was used for the first time to compare vectors of human similarity judgements in the pioneering study on the correlation of the degree of synonymy between words by Rubenstein and Goodenough (1965), which introduces the pioneering and most significant benchmark on word similarity. Pearson correlation has been subsequently used to create new word similarity and relatedness benchmarks, such as those detailed in Table 3, as well as for the evaluation of any method for the estimation of the degree of similarity between words and concepts. For this reason, we include the Pearson correlation values obtained by the methods evaluated herein with the aim of encouraging their comparison with most results reported in the literature. The Pearson correlation function returns a value that matches the normalised dot product between the two vectors representing the centred-samples of two random variables, which corresponds to the cosine function of the Euclidean angle between both vectors. Thus, Pearson correlation is invariant as regard any scaling or translation of the data, whilst the Spearman rank correlation is rank invariant what means that it holds the same value for any arbitrary monotone data transformation. On the other hand, Pearson correlation sets a linear and very strong form of correlation which makes it difficult their use as predictor of the impact of any word similarity method into some NLP and IR applications whose output are based on the ranking of different types of text-based information units according to their degree of similarity as regard any text-based query, such as document or sentence retrieval. Likewise, it is very helpful to be able to compare the intrinsic capability of any word similarity measure to rank correctly the degree of similarity between words or concepts regardless any arbitrary non-linear transformation of their output values. For these later reasons,

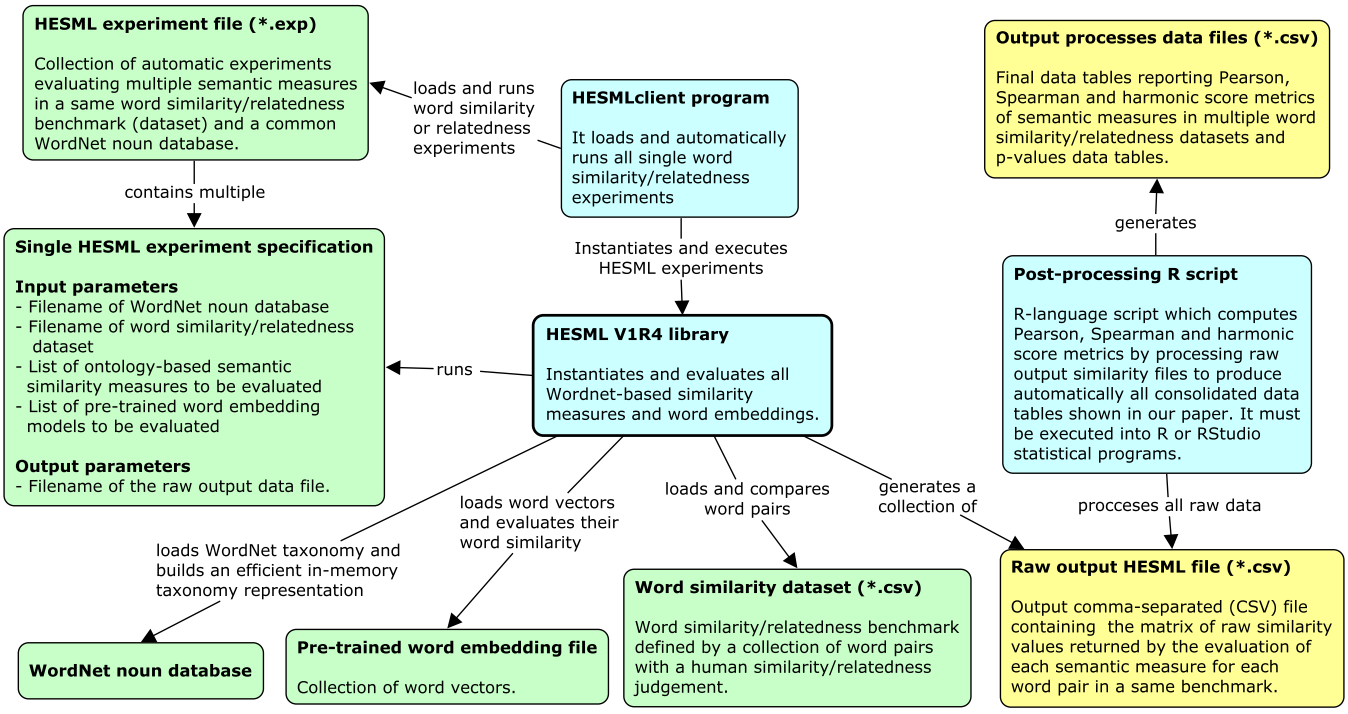

Fig. 3. Concept map detailing our experimental setup to run automatically all experiments reported herein. Input data files are shown in green, whilst output raw and processed data files are shown in yellow and software components are shown in blue. All experiments are specified into a single experiment file which is executed by HESMLclient program. For more detailed information, we refer the reader to Section 4.2 and Appendix B. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Collection of twenty-one ontology-based measures based on WordNet 3.0 which are evaluated in our experiments using HESML V1R4 (Lastra-Díaz and García Serrano, 2018).

Subfamily	Ontology-based similarity measures evaluated herein
Classic IC-based & reformulations	(1) Resnik (1995), (2) Lin (1998), (3) Jiang and Conrath (1997), (4) Pirró and Seco (2008)
Hybrid IC-based	(5) Cai et al. (2017, strategy 2), (6) (Lastra-Díaz and García-Serrano, 2015a, coswJ&C), (7) Zhou et al. (2008b), (8) Gao et al. (2015, strategy 3), (9) Meng et al. (2014)
Monotone IC-based transformations	(10) (Lastra-Díaz and García-Serrano, 2015a, cosJ&C), (11) Meng and Gu (2012), (12) FaITH (Pirró and Euzenat, 2010), (13) Garla and Brandt (2012)
Path-based	(14) Li et al. (2003, strategy 3), (15) Al-Mubaid and Nguyen (2009), (16) Pedersen et al. (2007), (17) (Leacock and Chodorow, 1998), (18) Rada et al. (1989), (19) Wu and Palmer (1994)
Feature-based	(20) Sánchez et al. (2012)
Taxonomy features	(21) Hadj Taieb et al. (2014b)

Table 2
Collection of eleven pre-trained Word Embedding (WE and WEC) models and Ontology-based Vector Models (OVM) evaluated in our experiments using the Java classes implementing their evaluation in HESML V1R4 (Lastra-Díaz and García Serrano, 2018). All pre-trained files are publicly available in a ZIP file within our reproducibility package provided as supplementary material (Lastra-Díaz et al., 2019). First column details which methods use WordNet.

Use WordNet	Family	Word Embedding or OVM model	Source of pre-trained model files
Yes	WEC	Attract-repel (Mrkšić et al., 2017)	https://github.com/nmrksic/attract-repel
No	WE	FastText (Bojanowski et al., 2016)	https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
No	WE	GloVe (Pennington et al., 2014)	https://nlp.stanford.edu/projects/glove/
No	WE	CBOW (Mikolov et al., 2013a)	https://code.google.com/archive/p/word2vec/
Yes	WEC	SymPatterns (SP-500d) (Schwartz et al., 2015)	https://homes.cs.washington.edu/~roysch/papers/sp_embeddings/sp_embeddings.html
No	WEC	Paragram-ws (Wieting et al., 2015)	https://www.cs.cmu.edu/~jwieting/
No	WEC	Paragram-sl (Wieting et al., 2015)	https://www.cs.cmu.edu/~jwieting/
Yes	WEC	Counter-fitting (CF) (Mrkšić et al., 2016)	https://github.com/nmrksic/counter-fitting
Yes	OVM	WN-RandomWalks (Goikoetxea et al., 2015)	http://ixa2.si.ehu.es/ukb/
Yes	OVM	WN-UKB (Agirre and Soroa, 2009)	http://ixa2.si.ehu.es/ukb/
Yes	OVM	Nasari (Camacho-Collados et al., 2016)	http://lcl.uniroma1.it/nasari/

the Spearman rank correlation has been widely adopted in this line of research with the hope that it can provide a more unbiased and intrinsic evaluation method and being a better predictor of the impact of any word similarity measure into rank-based applications. Finally, harmonic score allows to compare word similarity methods by using a unique weighted score based on their performance in Pearson and Spearman correlation.

In order to compare the performance of the semantic measures evaluated in our experiments, we use the average values of the three aforementioned metrics in all datasets. On the other hand, the statistical

Table 3
Detail of the main features of all datasets evaluated in our benchmarks, including the corresponding file in the companion reproducibility package (Lastra-Díaz et al., 2019) based on HESML V1R4 (Lastra-Díaz et al., 2017) and Reprozip (Chirigati et al., 2016).

Dataset	Content	Type	# word pairs	Filename (*.csv) in HESML distribution
MC28 (Miller and Charles, 1991)	Nouns	Similarity	28	Miller_Charles_28_dataset
RG65 (Rubenstein and Goodenough, 1965)	Nouns	Similarity	65	Rubenstein_Goodenough_dataset
PS_{full} (Pirró, 2009)	Nouns	Similarity	65	PirroSeco_full_dataset
Agirre201 (Agirre et al., 2009)	Nouns	Similarity	201	Agirre201_lowercase_dataset
SimLex665 (Hill et al., 2015)	Nouns	Similarity	665	SimLex665_dataset
MTurk771 (Halawi et al., 2012)	Nouns	Relatedness	771	Halawi_MTURK771_dataset
MTurk287/235 (Radinsky et al., 2011)	Nouns	Relatedness	235	Radinsky_MTurk287_filtered235_dataset
WS353Rel (Finkelstein et al., 2002)	Nouns	Relatedness	245	WordSim353Rel_dataset
Rel122 (Szumlanski et al., 2013)	Nouns	Relatedness	122	Rel122_dataset
SCWS (Huang et al., 2012)	Nouns	Relatedness	1994	SCWS1994_dataset
SimLex222 (Hill et al., 2015)	Verbs	Similarity	222	SimLex222_verbs_dataset
SimVerb3500 (Gerz et al., 2016)	Verbs	Similarity	3500	Gerz_SimVerb3500_dataset
YP130 (Yang and Powers, 2006)	Verbs	Relatedness	130	Yang_YP130_dataset
WS353Full (Finkelstein et al., 2002)	Nouns, Verbs, Adjectives	Relatedness	353	WordSim353Full_dataset
SimLex999 (Hill et al., 2015)	Nouns, Verbs, Adjectives	Similarity	999	SimLex999_dataset
MEN (Bruni et al., 2014)	Nouns, Verbs, Adjectives	Relatedness	3000	MEN_dataset
RW2034 (Luong et al., 2013)	Nouns, Verbs, Adjectives	Relatedness	2034	RareWords2034_dataset
RW1401 (Luong et al., 2013)	Nouns, Verbs, Adjectives in WordNet	Relatedness	2034	RareWords1401_dataset
SimLex111 (Hill et al., 2015)	Adjectives	Similarity	111	SimLex111_adjectives_dataset

significance of the results is evaluated by using the p-values resulting from the t-student test for the difference mean between the values reported by each pair of semantic measures in all datasets, or a subset of them relevant in the context of a discussion. The t-student test is used herein because it is a standard and widely-used hypothesis testing for small and independent data samples with normal distribution, in addition to being previously used in other similar studies by Lastra-Díaz and García-Serrano (2015a, 2016). The p-values are computed by using a one-sided t-student distribution on two paired random sample sets. Our null hypothesis, denoted by H_0, is that the difference in the average performance between each pair of compared semantic measures is 0, whilst the alternative hypothesis, denoted by H_1, is that their average performance is different. For a 5% level of significance, it means that if the p-value is greater or equal than 0.05, we must accept the null hypothesis, otherwise we can reject H_0 with an error probability of less than the p-value. In this latter case, we will said that a first semantic measure obtains a statistically significant higher value than a second one in a specific metric, or that the former one significantly outperforms the second one. All aforementioned metrics, as well as all p-values and data tables reported herein are computed by executing a R script file on the collection of raw similarity files generated by our experiments as detailed in Appendix B. Finally, in some cases we also complete our quantitative analysis with two qualitative judgments whose definition is as follows: (1) we say that a method is a *convincing winner* if it obtains the best average results and significantly outperforms most methods, with only a few exceptions, whilst (2) we say that a method is a *definitive winner* if it significantly outperforms the rest of methods. All p-values reported in this work are computed by calling the *t.test* function provided by the R package as detailed below:

t.test(tested method[], other method[], paired=TRUE, alternative ="greater")

4.4. Results obtained

Tables 4 and 5 show the Pearson and Spearman correlation values obtained by all semantic measures evaluated in the set of five noun similarity datasets made up by the MC28, RG65, $P\&S_{full}$, Agirre201 and SimLex665 datasets. Tables 6 and 7 show the performance obtained by all semantic measures in the set of four noun relatedness datasets made up by the MTurk771, $MTurk287_{235}$, WS353Rel and Rel122 datasets. Likewise, Tables 4 to 7 show the p-values testing the hypothesis that best performing methods in these benchmarks, Attract–repel and Paragram-ws models, significantly outperform the remaining methods. On the other hand, Tables 8 and 9 show the performance of all OVM and WE models in all similarity and relatedness datasets respectively, whose statistical significance analysis is shown in Tables A.1 and A.2 of Appendix A. In order to evaluate our hypothesis on the linear combination of two methods, Appendix A introduces Tables A.3 and A.4 showing the Pearson and Spearman correlation obtained by the combination of the best performing similarity measure with all remaining methods in the five aforementioned noun similarity datasets, whilst Tables A.5 and A.6 show the same aforementioned metrics obtained by the combination of the best performing relatedness measure with all remaining methods in four aforementioned noun relatedness datasets.

5. Discussion

5.1. WE models versus OM measures in noun similarity datasets

Attract–repel model obtains the highest average Pearson and Spearman correlation values in all noun similarity datasets. However, this model is unable to outperform significantly all ontology-based measures in any metric. This conclusion can be drawn by looking at the average (Avg) column in Tables 4 and 5 and p-values in last column of Table 4 which show that there is no a statistical significant difference between the Pearson correlation values obtained by the Attract–repel model and several ontology-based semantic similarity measures such as Cai et al. (p-value = 0.32) and coswJ&C (Lastra-Díaz and García-Serrano, 2015b) (p-value = 0.32) among others. Likewise, looking at p-values in last column of Table 5 we can see that Attract–repel model neither is able to outperform significantly in Spearman correlation all OTM measures, such as coswJ&C (p-value = 0.072), nor all OVM measures such as WN-RandomWalks (p-value = 0.27) and WordNet-UKB (p-value = 0.23). Thus, this conclusion allows to answer negatively the part (a) of research question RQ1 as follows: *the family of ontology-based semantic similarity measures has not been definitively outperformed by current state-of-the-art WE models in the estimation of the degree of similarity between words*.

CoswJ&C measure obtains the highest Pearson correlation values in MC28 and RG65 datasets, whilst Hadj Taieb et al. measure and GloVe and Counter-fitting models obtain the highest Pearson correlation values in $P\&S_{full}$, Agirre201 and SimLex-665 datasets respectively. This later conclusion can be drawn by looking at bold values in Table 4.

WN-RandomWalks, WordNet-UKB, Attract–repel, Paragram-ws and Counter-fitting obtain the highest Spearman rank correlation values in MC28, RG65, $P\&S_{full}$, Agirre201 and SimLex-665 datasets respectively. This later conclusion can be drawn by looking at bold values in Table 5.

Table 4
Pearson correlation (r) values for each Ontology-based semantic Topological similarity Measure (OTM), Ontology-based Vector Model (OVM) or Word Embedding (WE/WEC) model in the five noun similarity datasets. Measures (rows) are ranked according to their average value shown in Avg column. Best value for each dataset is shown in bold.

			Pearson correlation (r) in noun similarity datasets						
Family	Measure	IC model	MC28	RG65	PS_{full}	Agirre201	SL665	Avg (r)	p-value[a]
WEC	Attract–repel (Mrkšić et al., 2017)		0.837	0.840	0.893	0.720	0.691	**0.796**	–
OTM	$Cai_{strategy2}$ (Cai et al., 2017)	Cai et al. (2017)	0.858	0.872	0.901	0.687	0.608	0.785	0.320
OTM	coswJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	**0.871**	**0.877**	0.885	0.695	0.592	0.784	0.320
OTM	Zhou et al. (2008b)	Seco et al. (2004)	0.854	0.873	0.895	0.672	0.624	0.784	0.270
OTM	Hadj Taieb et al. (2014b)		0.825	0.867	**0.907**	0.708	0.609	0.783	0.260
OTM	cosJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.848	0.875	0.900	0.682	0.594	0.780	0.260
WEC	Paragram-ws (Wieting et al., 2015)		0.796	0.810	0.849	0.765	0.662	0.776	0.150
OTM	(Pirró and Seco, 2008)	Seco et al. (2004)	0.846	0.862	0.897	0.679	0.597	0.776	0.200
OTM	$Gao_{strategy3}$ (Gao et al., 2015)	CPRefHypo (Lastra-Díaz and García-Serrano, 2016)	0.835	0.865	0.891	0.674	0.614	0.776	0.160
WEC	Counter-fitting (Mrkšić et al., 2016)		0.806	0.806	0.866	0.701	**0.697**	0.775	0.022
OTM	Meng and Gu (2012)	Seco et al. (2004)	0.814	0.860	0.903	0.692	0.605	0.775	0.160
OTM	FaITH (Pirró and Euzenat, 2010)	Seco et al. (2004)	0.809	0.856	0.904	0.692	0.605	0.773	0.140
WEC	Paragram-sl (Wieting et al., 2015)		0.781	0.798	0.854	0.748	0.682	0.773	0.098
OTM	Lin (1998)	Seco et al. (2004)	0.824	0.861	0.894	0.680	0.601	0.772	0.140
OTM	$Li_{strategy3}$ (Li et al., 2003)		0.836	0.862	0.885	0.664	0.606	0.771	0.130
OTM	Jiang and Conrath (1997)	Sánchez et al. (2011)	0.859	0.862	0.876	0.652	0.584	0.767	0.160
OTM	Leacock and Chodorow (1998)		0.826	0.851	0.871	0.647	0.605	0.760	0.060
OVM	WN-RandowWalks (Goikoetxea et al., 2015)		0.835	0.797	0.843	0.773	0.543	0.758	0.160
OTM	Sánchez et al. (2012)		0.806	0.848	0.870	0.669	0.594	0.757	0.044
OTM	Meng et al. (2014)	Seco et al. (2004)	0.811	0.849	0.837	0.613	0.571	0.736	0.033
OTM	Al-Mubaid and Nguyen (2009)		0.791	0.807	0.853	0.645	0.576	0.734	0.008
OTM	Resnik (1995)	CPRefLeSubRat (Lastra-Díaz and García-Serrano, 2016)	0.793	0.823	0.874	0.669	0.512	0.734	0.054
WE	FastText (Bojanowski et al., 2016)		0.842	0.793	0.818	0.775	0.411	0.728	0.150
WE	GloVe (Pennington et al., 2014)		0.845	0.770	0.759	**0.797**	0.467	0.728	0.130
OVM	Nasari (Camacho-Collados et al., 2016)		0.831	0.791	0.812	0.708	0.489	0.726	0.060
WE	CBOW (Mikolov et al., 2013a)		0.796	0.772	0.786	0.763	0.461	0.716	0.073
OTM	Pedersen et al. (2007)		0.758	0.781	0.840	0.605	0.551	0.707	0.003
OTM	Garla and Brandt (2012)	Sánchez et al. (2011)	0.720	0.769	0.847	0.572	0.512	0.684	0.005
OTM	Rada et al. (1989)		0.729	0.771	0.751	0.558	0.565	0.675	0.001
OTM	$Wu\&Palmer_{fast}$(Wu and Palmer, 1994)		0.664	0.720	0.715	0.568	0.473	0.628	0.000
WEC	SymPatterns-500d (Schwartz et al., 2015)		0.606	0.690	0.709	0.454	0.435	0.579	0.000
OVM	WordNet UKB (Agirre et al., 2009)		0.542	0.548	0.629	0.375	0.361	0.491	0.000

[a]Last column shows p-values for an one-side t-Student distribution between Attract–repel (Mrkšić et al., 2017) model and the remaining methods using the performance in the five noun similarity datasets as paired random sample with the aim of testing the hypothesis that Attract–repel significantly outperforms remaining methods in Pearson correlation.

5.2. Comparison of WE models in noun similarity datasets

Attract–repel model obtains the best overall performance in all noun similarity datasets. However, it is unable to outperform significantly most WE models in any metric. This later conclusion can be drawn by looking at Pearson and Spearman correlation values in Tables 4 and 5, as well as p-values shown in these later tables. Thus, we conclude that there is no a definitive winner for the noun similarity task in the current family of WE models.

Attract–repel model significantly outperforms several WE and OVM models in all noun similarity datasets. First, looking at last column of Table 4, we can conclude that Attract–repel significantly outperforms in Pearson correlation the Counter-fitting (p-value = 0.022), SymPatterns-500d (p-value = 0.000), and WordNet-UKB (p-value = 0.000) models. And second, looking at last column of Table 5, we can conclude that Attract–repel significantly outperforms in Spearman correlation the Nasari (p-value = 0.008) and SymPatterns-500d (p-value = 0.000) models.

5.3. Comparison of OM measures in noun similarity datasets

WN-RandomWalks obtains the highest average Spearman correlation among the family of OM measures in all similarity datasets. However, there is no a statistical significant difference with most of best performing OM measures. The outperformance of WN-RandomWalks can be confirmed by looking at Table 5, whilst a t-Student significance analysis shows that there is no a statistical significance difference between the Spearman correlation values obtained by the WN-RandomWalks model and other OM measures, such as the coswJ&C (p-value = 0.329), Cai et al. (p-value = 0.203) or WordNet-UKB (p-value = 0.399) measures among others.

CoswJ&C similarity measure obtains the highest average Spearman correlation among the family of OTM measures in all noun similarity datasets. However, it is unable to outperform significantly all OTM measures. The outperformance of the *coswJ&C* measure can be confirmed by looking at Table 5, whilst a t-Student significance analysis shows that there is no a statistically significant difference between the Spearman correlation values obtained by the *coswJ&C* measure and those obtained by the Zhou et al. (p-value = 0.215), Cai et al. (p-value = 0.07028) and Meng et al. (p-value = 0.1733) measures. Thus, there is no a definitive winner in this aforementioned family of measures.

Cai et al. (2017) similarity measure obtains the highest average Pearson correlation value in all similarity datasets among the family of OTM measures. However, it is unable to outperform significantly all OTM measures. The outperformance of the Cai et al. measure can be confirmed by looking at Table 4, whilst a t-Student significance analysis shows that there is no a statistical significant difference with the Pearson correlation values obtained by the coswJ&C measure (p-value = 0.4108), and other measures such as those introduced by Zhou et al. (p-value = 0.3721), Hadj Taieb et al. (p-value = 0.4132) and cosJ&C (Lastra-Díaz and García-Serrano, 2015b) (p-value = 0.06963) among others.

5.4. WE models versus OM measures in noun relatedness datasets

GloVe model obtains the highest average Pearson correlation value in all noun relatedness datasets, whilst Paragram-ws obtains the highest average Spearman correlation value. Last conclusions can be drawn by looking at bold values (r = 0.679) and (ρ = 0.689) shown in first row of Tables 6 and 7 respectively.

GloVe model obtains the highest Pearson correlation value in MTurk771, $MTurk287_{235}$ and Rel122 datasets, whilst Paragram-ws model obtains the highest value in WS353Rel dataset. This later conclusion can be drawn by looking at bold values in Table 6.

Table 5
Spearman rank correlation (ρ) values for each Ontology-based semantic Topological similarity Measure (OTM), Ontology-based Vector Model (OVM) or Word Embedding (WE/WEC) model in the five noun similarity datasets. Measures (rows) are ranked according to their average value shown in Avg column. Best value for each dataset is shown in bold.

			Spearman correlation (ρ) in noun similarity datasets						
Family	Measure	IC model	MC28	RG65	PS_{full}	Agirre201	SL665	Avg (ρ)	p-value[a]
WEC	Attract–repel (Mrkšić et al., 2017)		0.884	0.825	**0.843**	0.738	0.690	**0.796**	–
WEC	Paragram-ws (Wieting et al., 2015)		0.824	0.813	0.821	**0.808**	0.645	0.782	0.290
WEC	Counter-fitting (Mrkšić et al., 2016)		0.857	0.808	0.831	0.695	**0.698**	0.778	0.050
OVM	WN-RandowWalks (Goikoetxea et al., 2015)		**0.909**	0.823	0.814	0.784	0.529	0.772	0.270
OVM	WordNet UKB (Agirre et al., 2009)		0.894	**0.858**	0.841	0.718	0.524	0.767	0.230
OTM	coswJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.877	0.835	0.822	0.666	0.587	0.757	0.072
WEC	Paragram-sl (Wieting et al., 2015)		0.761	0.775	0.794	0.778	0.679	0.757	0.110
OTM	Zhou et al. (2008b)	Seco et al. (2004)	0.846	0.824	0.814	0.655	0.610	0.750	0.022
OTM	$Cai_{strategy2}$ (Cai et al., 2017)	Cai et al. (2017)	0.864	0.804	0.794	0.662	0.595	0.744	0.012
OTM	Meng et al. (2014)	Seco et al. (2004)	0.805	0.820	0.815	0.655	0.610	0.741	0.014
OTM	Pirró and Seco (2008)	Seco et al. (2004)	0.868	0.801	0.792	0.656	0.586	0.740	0.015
OTM	cosJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.849	0.803	0.800	0.650	0.591	0.739	0.010
OTM	Jiang and Conrath (1997)	Sánchez et al. (2011)	0.849	0.803	0.800	0.650	0.591	0.739	0.010
OTM	Garla and Brandt (2012)	Sánchez et al. (2011)	0.849	0.803	0.800	0.650	0.591	0.739	0.010
OTM	Hadj Taieb et al. (2014b)		0.807	0.797	0.797	0.660	0.596	0.731	0.003
OTM	$Gao_{strategy3}$ (Gao et al., 2015)	CPRefHypo (Lastra-Díaz and García-Serrano, 2016)	0.827	0.801	0.791	0.641	0.595	0.731	0.005
OTM	Meng and Gu (2012)	Seco et al. (2004)	0.831	0.797	0.791	0.647	0.589	0.731	0.004
OTM	FaITH (Pirró and Euzenat, 2010)	Seco et al. (2004)	0.831	0.797	0.791	0.647	0.589	0.731	0.004
OTM	Lin (1998)	Seco et al. (2004)	0.831	0.797	0.791	0.647	0.589	0.731	0.004
OTM	Al-Mubaid and Nguyen (2009)		0.805	0.812	0.807	0.645	0.578	0.729	0.011
WE	FastText (Bojanowski et al., 2016)		0.845	0.801	0.801	0.777	0.410	0.727	0.140
OTM	$Li_{strategy3}$ (Li et al., 2003)		0.813	0.810	0.798	0.625	0.588	0.727	0.009
OTM	Pedersen et al. (2007)		0.813	0.810	0.798	0.625	0.588	0.727	0.009
OTM	Leacock and Chodorow (1998)		0.813	0.810	0.798	0.625	0.588	0.727	0.009
OTM	Rada et al. (1989)		0.813	0.810	0.798	0.625	0.588	0.727	0.009
WE	GloVe (Pennington et al., 2014)		0.862	0.769	0.755	0.795	0.429	0.722	0.120
OTM	Sánchez et al. (2012)		0.790	0.784	0.789	0.643	0.578	0.717	0.002
WE	CBOW (Mikolov et al., 2013a)		0.781	0.760	0.767	0.772	0.454	0.707	0.055
OTM	Resnik (1995)	CPRefLeSubRat Lastra-Díaz and García-Serrano (2016)	0.839	0.763	0.757	0.638	0.511	0.702	0.008
OVM	Nasari (Camacho-Collados et al., 2016)		0.796	0.745	0.752	0.684	0.488	0.693	0.008
OTM	$Wu\&Palmer_{fast}$(Wu and Palmer, 1994)		0.602	0.712	0.716	0.600	0.482	0.623	0.003
WEC	SymPatterns-500d (Schwartz et al., 2015)		0.652	0.663	0.674	0.483	0.460	0.587	0.000

[a]Last column shows p-values for an one-side t-Student distribution between Attract–repel (Mrkšić et al., 2017) model and the remaining methods using the performance in the five noun similarity datasets as paired random sample with the aim of testing the hypothesis that Attract–repel significantly outperforms remaining methods in Spearman rank correlation.

Paragram-ws model obtains the highest Spearman correlation value in MTurk771, WS353Rel and Rel122 datasets, whilst GloVe model obtains the highest value in $MTurk287_{235}$ dataset. This later conclusion can be drawn by looking at bold values in Table 7.

All WE models obtain much higher average Pearson and Spearman correlation values than all ontology-based semantic similarity measures in all noun relatedness datasets. Looking at Tables 6 and 7, we can conclude that all WE models are able to outperform all ontology-based semantic similarity measures in all aforementioned metrics. This conclusion confirms a well-known fact in the research community on the outperformance of corpus-based methods on the ontology-based measures based on WN in the estimation of the degree of relatedness between words and concepts. Ontology-based measures based on WN only use 'is-a' relationships, thus they are unable to capture semantic relationships between words and concepts as done by the family of WE models which identify their co-occurrence in text.

Paragram-ws model significantly outperforms all ontology-based semantic similarity measures based on WN in all metrics and all noun relatedness datasets. Looking at last columns in Tables 6 and 7, we can see that all p-values obtained for the statistical significance analysis on the Pearson and Spearman correlation values obtained by the Paragram-ws model as regard to the rest of methods in all noun relatedness datasets are lower than 0.05 for all ontology-based semantic similarity measures. Thus, this conclusion allows to answer positively the part (b) of research question RQ1 as follows: *the family of ontology-based semantic similarity measures has been definitively outperformed by current state-of-the-art WE models in the estimation of the degree of relatedness between words*.

5.5. Comparison of WE models in noun relatedness datasets

Paragram-ws model significantly outperforms all WE models in Pearson and Spearman correlation in all noun relatedness datasets, with the only exception of the GloVe, FastText and CBOW models. Looking at average (Avg) column in Tables 6 and 7, we can see that Paragram-ws model obtains the second highest average Pearson correlation, just behind GloVe model, and the highest average Spearman correlation value. Likewise, looking at p-values in Tables 6 and 7, we can see that all p-values comparing the performance in Pearson and Spearman correlation between Paragram-ws model and the rest of WE models are lower than 0.05, with the only exception of those obtained by the GloVe, FastText and CBOW models. Thus, Paragram-ws model is a convincing winner for the noun relatedness task among the family of WE models.

5.6. Comparison of OM measures in noun relatedness datasets

Hadj Taieb et al. measure obtains the highest average Pearson correlation value in all noun relatedness datasets among the family of OTM measures. However, there is no a statistical significant difference among them. This conclusion can be drawn by looking at average column in Table 6 and checking the p-values resulting from the t-Student test comparing the values reported by the Hadj Taieb et al. measure and other measures such as coswJ&C (p-value = 0.4785), $Cai_{strategy2}$ (p-value = 0.4758) and FaITH (p-value = 0.3539) measures.

CoswJ&C measure obtains the highest average Spearman correlation values in all noun relatedness datasets among the family of OTM measures. However, there is no a statistical significant difference among them. This conclusion can be drawn by looking at average (Avg) column Table 7 and checking the p-values resulting from the t-Student test comparing the values reported by the coswJ&C measure and other measures such

Table 6
Pearson correlation (r) values for each Ontology-based semantic Topological similarity Measure (OTM), Ontology-based Vector Model (OVM) or Word Embedding (WE/WEC) model in the four noun relatedness datasets. Measures (rows) are ranked according to their average value shown in Avg column. Best value for each dataset is shown in bold.

			Pearson correlation (r) in noun relatedness datasets					
Family	Measure	IC model	MTurk771	MTurk287$_{235}$	WS353Rel	Rel122	Avg (r)	p-value[a]
WE	GloVe (Pennington et al., 2014)		**0.705**	**0.749**	0.665	**0.599**	**0.679**	0.900
WEC	Paragram-ws (Wieting et al., 2015)		0.701	0.704	**0.668**	0.573	0.661	–
WE	FastText (Bojanowski et al., 2016)		0.641	0.728	0.659	0.540	0.642	0.180
WE	CBOW (Mikolov et al., 2013a)		0.650	0.698	0.567	0.576	0.623	0.100
WEC	Paragram-sl (Wieting et al., 2015)		0.674	0.660	0.578	0.523	0.609	0.014
OVM	WN-RandowWalks (Goikoetxea et al., 2015)		0.642	0.653	0.623	0.513	0.608	0.000
WEC	Counter-fitting (Mrkšić et al., 2016)		0.666	0.630	0.590	0.526	0.603	0.006
WEC	Attract–repel (Mrkšić et al., 2017)		0.590	0.618	0.541	0.464	0.553	0.000
OVM	Nasari (Camacho-Collados et al., 2016)		0.505	0.474	0.400	0.553	0.483	0.024
WEC	SymPatterns-500d (Schwartz et al., 2015)		0.506	0.409	0.328	0.436	0.420	0.007
OVM	WordNet UKB (Agirre et al., 2009)		0.349	0.320	0.414	0.328	0.353	0.002
OTM	Hadj Taieb et al. (2014b)		0.540	0.496	0.091	0.189	0.329	0.019
OTM	coswJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.529	0.487	0.079	0.220	0.328	0.019
OTM	Cai$_{strategy2}$ (Cai et al., 2017)	Cai et al. (2017)	0.536	0.478	0.081	0.219	0.328	0.019
OTM	FaITH (Pirró and Euzenat, 2010)	Seco et al. (2004)	0.515	0.490	0.088	0.208	0.325	0.017
OTM	Meng and Gu (2012)	Seco et al. (2004)	0.520	0.487	0.084	0.204	0.324	0.017
OTM	Pirró and Seco (2008)	Seco et al. (2004)	0.515	0.482	0.085	0.210	0.323	0.016
OTM	Zhou et al. (2008b)	Seco et al. (2004)	0.541	0.467	0.069	0.211	0.322	0.019
OTM	cosJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.525	0.480	0.079	0.188	0.318	0.017
OTM	Lin (1998)	Seco et al. (2004)	0.524	0.469	0.069	0.193	0.314	0.017
OTM	Gao$_{strategy3}$ (Gao et al., 2015)	CPRefHypo (Lastra-Díaz and García-Serrano, 2016)	0.533	0.454	0.072	0.184	0.311	0.017
OTM	Jiang and Conrath (1997)	Sánchez et al. (2011)	0.518	0.442	0.066	0.194	0.305	0.015
OTM	Li$_{strategy3}$ (Li et al., 2003)		0.518	0.434	0.070	0.190	0.303	0.014
OTM	Garla and Brandt (2012)	Sánchez et al. (2011)	0.410	0.437	0.117	0.234	0.300	0.006
OTM	Pedersen et al. (2007)		0.425	0.423	0.103	0.221	0.293	0.006
OTM	Leacock and Chodorow (1998)		0.506	0.410	0.056	0.195	0.292	0.013
OTM	Resnik (1995)	CPRefLeSubRat (Lastra-Díaz and García-Serrano, 2016)	0.402	0.453	0.077	0.195	0.282	0.007
OTM	Al-Mubaid and Nguyen (2009)		0.446	0.415	0.059	0.204	0.281	0.009
OTM	Sánchez et al. (2012)		0.506	0.394	0.057	0.160	0.279	0.011
OTM	Meng et al. (2014)	Seco et al. (2004)	0.471	0.425	0.032	0.161	0.272	0.012
OTM	Rada et al. (1989)		0.510	0.338	0.020	0.175	0.261	0.012
OTM	Wu&Palmer$_{fast}$ (Wu and Palmer, 1994)		0.466	0.302	0.024	0.159	0.238	0.008

[a]Last column shows p-values for an one-side t-Student distribution between Paragram-ws (Wieting et al., 2015) model and the remaining methods using the performance in the four noun relatedness datasets as paired random sample with the aim of testing the hypothesis that Paragram-ws significantly outperforms remaining methods in Pearson correlation.

as Cai$_{strategy2}$ (p-value = 0.3846) and Zhou et al. (p-value = 0.3555) measures among others

WN-RandomWalks outperforms the Nasari and WordNet-UKB OVM models in average Pearson correlation in all noun relatedness datasets. However, there is no a statistical significant difference with Nasari. This conclusion can be drawn by looking at average (Avg) column in Table 6 and checking p-value resulting from a t-Student test comparing the values reported by WN-RandomWalks and WordNet-UKB (p-value = 0.003) and Nasari (p-value = 0.0596).

WN-RandomWalks outperforms the WordNet-UKB and Nasari OVM models in average Spearman correlation in all noun relatedness datasets. However, there is no a statistical significant difference among them. Looking at average (Avg) column in Table 7, we can see that WN-RandomWalks outperforms in average Spearman correlation both WordNet-UKB and Nasari models. However, a t-Student significance analysis shows that there is no a statistical significant difference in Spearman correlation between WN-RandomWalks and WordNet-UKB (p-value = 0.065) or Nasari (p-value = 0.051) models. Thus, WN-RandomWalks is a convincing but not definitive winner among the family of OVM models in all noun relatedness datasets.

WN-RandomWalks significantly outperforms in all metrics all OTM measures in all noun relatedness datasets. Looking at Tables 6 and 7, we can see that WN-RandomWalks obtains average Pearson and Spearman correlation values of (r=0.608) and (ρ=0.615) respectively, whilst the best performing OTM measures obtain values of (r=0.329) and (ρ=0.280) respectively. Likewise, a t-Student test shows that this difference is statistically significant.

5.7. Comparison of WE models in all similarity datasets

Attract–repel model obtains the highest average Pearson, Spearman and harmonic values in all similarity datasets. This conclusion can be drawn by looking at average values in Table 8. Likewise, Attract–repel model obtains the best results for the aforementioned metrics in most of similarity datasets, with only a few exceptions as shown by bold values in Table 8.

Attract–repel model significantly outperforms the rest of WE models in all metrics and all similarity datasets. This conclusion can be drawn by looking at Tables A.1 and checking that all p-values for WE models in all metrics are lower than 0.05. Thus, Attract–repel model is a definitive winner for the estimation of the degree of similarity between words among the family of WE models.

5.8. Comparison of WE models in all relatedness datasets

Paragram-ws model obtains the highest average Pearson, Spearman and harmonic values in all relatedness datasets. This conclusion can be drawn by looking at average values in Table 9. Likewise, Paragram-ws model obtains the best Spearman correlation and harmonic score in most of relatedness datasets, with only a few exceptions as shown by bold values in Table 9.

Paragram-ws model significantly outperforms the rest of WE models in all metrics and all relatedness datasets, with the only exception of GloVe model. This conclusion can be drawn by looking at Tables A.2 and checking that all p-values for WE models in all metrics are lower than 0.05, with the only exception of those p-values shown by GloVe in Pearson correlation (0.23) and harmonic score (0.072). Thus, Paragram-ws model is a convincing winner for the estimation of the degree of relatedness between words among the family of WE models.

GloVe and Paragram-ws are the two best performing WE models in Pearson correlation in all relatedness datasets. Looking at Table 9, we can see that Glove model obtains the highest Pearson correlation value in MTurk771, MTurk287$_{235}$, Rel122, WS353Full and MEN datasets, whilst

Table 7
Spearman rank correlation (ρ) values for each Ontology-based semantic Topological similarity Measure (OTM), Ontology-based Vector Model (OVM) or Word Embedding (WE/WEC) model in the four noun relatedness datasets. Measures (rows) are ranked according to the average value shown in Avg column. Best value for each dataset is shown in bold.

			Spearman correlation (ρ) in noun relatedness datasets					
Family	Measure	IC model	MTurk771	MTurk287$_{235}$	WS353Rel	Rel122	Avg (ρ)	p-value[a]
WEC	Paragram-ws (Wieting et al., 2015)		**0.745**	0.699	**0.721**	**0.589**	**0.689**	–
WE	GloVe (Pennington et al., 2014)		0.715	**0.724**	0.651	0.584	0.669	0.200
WE	FastText (Bojanowski et al., 2016)		0.661	0.709	0.681	0.539	0.648	0.062
WEC	Paragram-sl (Wieting et al., 2015)		0.712	0.663	0.644	0.560	0.645	0.015
WEC	Counter-fitting (Mrkšić et al., 2016)		0.701	0.639	0.645	0.562	0.637	0.008
WE	CBOW (Mikolov et al., 2013a)		0.672	0.674	0.603	0.560	0.627	0.034
OVM	WN-RandowWalks (Goikoetxea et al., 2015)		0.672	0.640	0.626	0.521	0.615	0.001
WEC	Attract–repel (Mrkšić et al., 2017)		0.599	0.606	0.586	0.466	0.564	0.001
OVM	WordNet UKB (Agirre et al., 2009)		0.602	0.599	0.488	0.525	0.553	0.017
OVM	Nasari (Camacho-Collados et al., 2016)		0.500	0.434	0.442	0.557	0.483	0.019
WEC	SymPatterns-500d (Schwartz et al., 2015)		0.513	0.414	0.303	0.405	0.409	0.006
OTM	coswJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.481	0.367	0.020	0.252	0.280	0.013
OTM	Cai$_{strategy2}$ (Cai et al., 2017)	Cai et al. (2017)	0.491	0.369	0.044	0.193	0.274	0.010
OTM	Zhou et al. (2008b)	Seco et al. (2004)	0.508	0.369	0.031	0.175	0.271	0.012
OTM	Pirró and Seco (2008)	Seco et al. (2004)	0.465	0.385	0.049	0.181	0.270	0.009
OTM	cosJ&C (Lastra-Díaz and García-Serrano, 2015b)	Sánchez et al. (2011)	0.479	0.365	0.037	0.189	0.267	0.010
OTM	Jiang and Conrath (1997)	Sánchez et al. (2011)	0.479	0.365	0.037	0.189	0.267	0.010
OTM	Garla and Brandt (2012)	Sánchez et al. (2011)	0.479	0.365	0.037	0.189	0.267	0.010
OTM	Hadj Taieb et al. (2014b)		0.501	0.337	0.032	0.174	0.261	0.010
OTM	Meng and Gu (2012)	Seco et al. (2004)	0.479	0.339	0.033	0.180	0.257	0.009
OTM	FaITH (Pirró and Euzenat, 2010)	Seco et al. (2004)	0.479	0.339	0.033	0.180	0.257	0.009
OTM	Lin (1998)	Seco et al. (2004)	0.479	0.339	0.033	0.180	0.257	0.009
OTM	Meng et al. (2014)	Seco et al. (2004)	0.503	0.331	0.007	0.167	0.252	0.011
OTM	Sánchez et al. (2012)		0.483	0.302	0.026	0.171	0.245	0.008
OTM	Gao$_{strategy3}$ (Gao et al., 2015)	CPRefHypo (Lastra-Díaz and García-Serrano, 2016)	0.486	0.303	0.004	0.181	0.244	0.010
OTM	Al-Mubaid and Nguyen (2009)		0.492	0.311	0.003	0.167	0.243	0.010
OTM	Li$_{strategy3}$ (Li et al., 2003)		0.490	0.311	−0.002	0.156	0.239	0.010
OTM	Pedersen et al. (2007)		0.490	0.311	−0.002	0.156	0.239	0.010
OTM	Leacock and Chodorow (1998)		0.490	0.311	−0.002	0.156	0.239	0.010
OTM	Rada et al. (1989)		0.490	0.311	−0.002	0.156	0.239	0.010
OTM	Resnik (1995)	CPRefLeSubRat Lastra-Díaz and García-Serrano (2016)	0.366	0.333	0.041	0.184	0.231	0.004
OTM	Wu&Palmer$_{fast}$(Wu and Palmer, 1994)		0.448	0.280	0.030	0.159	0.229	0.006

[a]Last column shows p-values for an one-side t-Student distribution between Paragram-ws (Wieting et al., 2015) model and the remaining methods using the performance in the four noun relatedness datasets as paired random sample with the aim of testing the hypothesis that Paragram-ws significantly outperforms remaining methods in Spearman rank correlation.

Paragram-ws obtains the highest value in WS353Rel, RW2034 and RW1401, WN-RandomWalks in YP130 and Paragram-sl in SCWS1994 datasets.

5.9. Impact of WordNet in WE models

Most of the embedding models evaluated in this survey which incorporate KB information use WN, among others. SymPatterns-500d (Schwartz et al., 2015) is based on antonymy relations of WN and a thesaurus, Counter-fitting makes use antonymy and synonymy relations from WN and PPDB, and Attract–repel antonymy and synonymy relations from BabelNet (Navigli and Ponzetto, 2012), which includes WN. The only exception is Paragram, which is based on PPDB paraphrases.

WE models that combine distributional representations with relational information from KBs have shown a robust performance in both similarity and relatedness measurements. Special mention goes to Attract–repel and Paragram-ws models, which respectively achieve state-of-the-art results in similarity (see Table 8) and relatedness (see Table 9) datasets evaluated herein. Attract–repel method uses synonymy and antonymy relations from several KBs within BabelNet, so that it is not surprising its state-of-the-art results in similarity datasets. However, Attract–repel shows weak results in relatedness datasets (see Table 9) due to its specialisation in similarity relations. Note that even though Paragram-ws is based on paraphrase relations, i.e. similarity relations, and despite the emphasis of the authors on similarity relations (see Section 3.2.2), it shows a robust performance in both similarity (see Table 8) and very specially in relatedness (see Table 9) within the WE models. Counter-fitting method also shows a robust performance in similarity datasets, so that it is very close to the state-of-the-art Attract–repel (see Table 8). However, its results in relatedness are well below the state-of-the-art Paragram-ws and also fall below the text-based word representations (see Table 9). SymmPatterns-500d shows very poor performance in both similarity (see Tables 4 and 5) and relatedness datasets by far (see Tables 6 and 7). Note that all ontology-based (OM) methods are based on WN taxonomy. As such, the methods that combine distributional representations with ontology information, share a large part of the structural information.

State-of-the-art Attract–repel and Paragram-ws outperform ontology-based measures in similarity and relatedness datasets respectively. Tables 4 and 5 show that the former methods outperform all ontology-based measures, but the difference is not significant. Paragram-ws slightly outperforms ontology-based measures in similarity (see Table 5), whilst it significantly outperforms OM measures in relatedness (see Tables 6 and 7). Counter-fitting's results are very close to the ontology-based ones in similarity (see Tables 4 and 5) and slightly outperforms them in relatedness (see Tables 6 and 7). SymmPatterns-500d results fall far below OM models.

The combination of WN, among other KBs, with distributional information has been key for word embedding methods to outperform OM in similarity. Distributional WE methods perform below OM methods in word similarity, while combined methods (WEC) are able to outperform OM methods in similarity. Although some WEC methods use WN in conjunction with other KBs like PPDB or as subsets of larger knowledge bases like BabelNet, the results indicate that WN is key to outperform OM and WE methods, although the difference is not statistically significant. Thus, this conclusion allows to answer positively to the version of research question RQ2 as follows: *The use of WN into recent WE models has been key to outperform previous OM and WE methods in the word similarity task, although the difference is not statistically significant*.

Table 8
Pearson (r), Spearman (ρ) and Harmonic score (h) metrics obtained by each Word Embedding (WE/WEC) or Ontology-based Vector (OVM) model in all similarity datasets. Word embedding (WE/WEC) and ontology-based vector models (columns) are ranked in descending order from left to right according to their average harmonic score shown in last row. Best value for each dataset and metric is shown in bold.

	Attract–repel (Mrkšić et al., 2017)	Counter-fitting (Mrkšić et al., 2016)	Paragram-ws (Wieting et al., 2015)	Paragram-ws (Wieting et al., 2015)	WN-RandowWalks (Goikoetxea et al., 2015)	CBOW (Mikolov et al., 2013a)	Nasari (Camacho-Collados et al., 2016)	GloVe (Pennington et al., 2014)	FastText (Bojanowski et al., 2016)	SymPatterns-500d (Schwartz et al., 2015)	WordNet UKB (Agirre et al., 2009)
Method and family	WEC	WEC	WEC	WEC	OVM	WE	OVM	WE	WE	WEC	OVM
Dataset	Pearson (r) correlation values in all similarity datasets										
MC28	0.837	0.806	0.796	0.781	0.835	0.796	0.831	**0.845**	0.842	0.606	0.542
RG65	**0.840**	0.806	0.810	0.798	0.797	0.772	0.791	0.770	0.793	0.690	0.548
PS_{full}	**0.893**	0.866	0.849	0.854	0.843	0.786	0.812	0.759	0.818	0.709	0.629
Agirre201	0.720	0.701	0.765	0.748	0.773	0.763	0.708	**0.797**	0.775	0.454	0.375
SimLex665	0.691	**0.697**	0.662	0.682	0.543	0.461	0.489	0.467	0.411	0.435	0.361
SimLex111	**0.877**	0.857	0.844	0.815	0.637	0.598	0.498	0.614	0.484	0.700	0.443
SimLex222	**0.777**	0.713	0.574	0.605	0.464	0.349	0.428	0.220	0.247	0.537	0.338
SimLex999	**0.745**	0.728	0.676	0.689	0.532	0.454	0.466	0.437	0.385	0.493	0.370
SimVerb3500	**0.666**	0.613	0.524	0.546	0.549	0.375	0.336	0.294	0.263	0.327	0.387
Avg (r)	**0.783**	0.754	0.722	0.724	0.664	0.595	0.595	0.578	0.558	0.550	0.444
Dataset	Spearman (ρ) correlation values in all similarity datasets										
MC28	0.884	0.857	0.824	0.761	**0.909**	0.781	0.796	0.862	0.845	0.652	0.894
RG65	0.825	0.808	0.813	0.775	0.823	0.760	0.745	0.769	0.801	0.663	**0.858**
PS_{full}	**0.843**	0.831	0.821	0.794	0.814	0.767	0.752	0.755	0.801	0.674	0.841
Agirre201	0.738	0.695	**0.808**	0.778	0.784	0.772	0.684	0.795	0.777	0.483	0.718
SimLex665	0.690	**0.698**	0.645	0.679	0.529	0.454	0.488	0.429	0.410	0.460	0.524
SimLex111	**0.872**	0.847	0.825	0.795	0.643	0.592	0.473	0.622	0.508	0.676	0.555
SimLex222	**0.783**	0.727	0.562	0.590	0.446	0.322	0.414	0.196	0.231	0.544	0.367
SimLex999	**0.751**	0.736	0.667	0.685	0.525	0.442	0.450	0.408	0.380	0.513	0.497
SimVerb3500	**0.672**	0.628	0.514	0.540	0.545	0.364	0.287	0.283	0.258	0.328	0.499
Avg (ρ)	**0.784**	0.758	0.720	0.711	0.669	0.584	0.566	0.569	0.557	0.555	0.639
Dataset	Harmonic score (h) values in all similarity datasets										
MC28	0.860	0.831	0.810	0.771	**0.870**	0.788	0.813	0.853	0.844	0.628	0.675
RG65	**0.833**	0.807	0.811	0.786	0.810	0.766	0.767	0.770	0.797	0.676	0.669
PS_{full}	**0.867**	0.848	0.835	0.823	0.828	0.777	0.781	0.757	0.809	0.691	0.720
Agirre201	0.729	0.698	0.786	0.762	0.779	0.767	0.696	**0.796**	0.776	0.468	0.493
SimLex665	0.690	**0.697**	0.653	0.681	0.536	0.457	0.489	0.447	0.411	0.447	0.427
SimLex111	**0.874**	0.852	0.835	0.805	0.640	0.595	0.485	0.618	0.496	0.688	0.493
SimLex222	**0.780**	0.720	0.568	0.597	0.455	0.335	0.421	0.207	0.239	0.540	0.352
SimLex999	**0.748**	0.732	0.671	0.687	0.528	0.448	0.458	0.422	0.383	0.503	0.424
SimVerb3500	**0.669**	0.620	0.519	0.543	0.547	0.369	0.310	0.288	0.261	0.328	0.436
Avg (h)	**0.783**	0.756	0.721	0.717	0.666	0.589	0.580	0.573	0.557	0.552	0.521

5.10. Impact of the averaging of WE models

Combinations of Attract–repel model with the measures by Zhou et al. and Cai et al. significantly outperform all state-of-the-art single methods in noun similarity datasets. This conclusion can be drawn by looking at results shown in Tables A.3 and A.4.

A large set of combinations of Attract–repel model with other OTM measures significantly outperform current state-of-the-art results in all noun similarity datasets. This conclusion can be drawn by comparing the results shown in Tables A.3 and A.4 with the best state-the-art results shown in Tables 4 and 5.

Combinations of Paragram-ws model with GloVe and WN-RandomWalks models significantly outperform all state-of-the-art single methods in noun relatedness datasets. This conclusion can be drawn by looking at results shown in Tables A.5 and A.6.

Combinations of Paragram-ws model with GloVe and WN-RandomWalks models significantly outperform current state-of-the-art results in all noun relatedness datasets. This conclusion can be drawn by comparing the results shown in Tables A.5 and A.6 with best state-the-art results shown in Tables 6 and 7.

All combinations of Attract–repel model with other measures obtain statistically significant higher correlation values than their corresponding base measures in noun similarity datasets with only two exceptions. Looking at Tables A.3 and A.4, we can see that all p-values are lower than 0.05 with the only exception of FastText model in Pearson correlation (p-value = 0.055) and WordNet-UKB model in Spearman correlation (p-value = 0.130) respectively.

All combinations of Paragram-ws model with other measures obtain statistically significant higher correlation values than their corresponding base measures in noun relatedness datasets. Looking at Tables A.3 and A.4, we can see that all p-values are lower than 0.05.

Any linear combination of the best performing similarity and relatedness measures significantly improves the performance of its base measure with only two exceptions. This conclusion follows from two previous conclusions above for the combination of the Attract–repel and Paragram-ws models. Thus, these findings allow to confirm positively RQ4 question.

Table 9
Pearson (r), Spearman (ρ) and Harmonic score (h) metrics obtained by each Word Embedding (WE/WEC) or Ontology-based Vector (OVM) model in all relatedness datasets. Word embedding (WE/WEC) and ontology-based vector models (columns) are ranked in descending order from left to right according to their average harmonic score shown in last row. Best value for each dataset and metric is shown in bold.

	Paragram-ws (Wieting et al., 2015)	Paragram-sl (Wieting et al., 2015)	GloVe (Pennington et al., 2014)	FastText (Bojanowski et al., 2016)	WN-RandowWalks (Goikoetxea et al., 2015)	CBOW (Mikolov et al., 2013a)	Counter-fitting (Mrkšić et al., 2016)	Attract–repel (Mrkšić et al., 2017)	Nasari (Camacho-Collados et al., 2016)	WordNet UKB (Agirre et al., 2009)	SymPatterns-500d (Schwartz et al., 2015)
Method and family	WEC	WEC	WE	WE	OVM	WE	WEC	WEC	OVM	OVM	WEC
Dataset	Pearson (r) correlation values in all relatedness datasets										
MTurk771	0.701	0.674	**0.705**	0.641	0.642	0.650	0.666	0.590	0.505	0.349	0.506
$MTurk287_{235}$	0.704	0.660	**0.749**	0.728	0.653	0.698	0.630	0.618	0.474	0.320	0.409
WS353Rel	**0.668**	0.578	0.665	0.659	0.623	0.567	0.590	0.541	0.400	0.414	0.328
Rel122	0.573	0.523	**0.599**	0.540	0.513	0.576	0.526	0.464	0.553	0.328	0.436
WS353Full	0.679	0.640	**0.713**	0.698	0.667	0.642	0.615	0.608	0.539	0.300	0.373
MEN	0.754	0.712	**0.800**	0.755	0.725	0.723	0.680	0.655	0.626	0.362	0.438
YP130	0.721	0.712	0.509	0.517	**0.792**	0.557	0.703	0.746	0.648	0.583	0.407
RW2034	**0.505**	0.498	0.440	0.432	0.253	0.438	0.288	0.319	0.139	0.269	0.229
RW1401	**0.518**	0.511	0.465	0.451	0.453	0.448	0.295	0.329	0.166	0.335	0.263
SCWS1994	0.115	**0.116**	0.106	0.106	0.110	0.105	0.114	0.113	0.084	0.085	0.089
Avg (r)	**0.594**	0.562	0.575	0.553	0.543	0.540	0.511	0.498	0.413	0.335	0.348
Dataset	Spearman (ρ) correlation values in all relatedness datasets										
MTurk771	**0.745**	0.712	0.715	0.661	0.672	0.672	0.701	0.599	0.500	0.602	0.513
$MTurk287_{235}$	0.699	0.663	**0.724**	0.709	0.640	0.674	0.639	0.606	0.434	0.599	0.414
WS353Rel	**0.721**	0.644	0.651	0.681	0.626	0.603	0.645	0.586	0.442	0.488	0.303
Rel122	**0.589**	0.560	0.584	0.539	0.521	0.560	0.562	0.466	0.557	0.525	0.405
WS353Full	**0.764**	0.716	0.716	0.738	0.718	0.684	0.680	0.666	0.567	0.606	0.391
MEN	0.799	0.771	**0.801**	0.762	0.754	0.732	0.741	0.709	0.639	0.669	0.434
YP130	0.669	0.655	0.535	0.509	**0.777**	0.530	0.621	0.655	0.568	0.686	0.361
RW2034	**0.536**	0.533	0.451	0.464	0.264	0.453	0.207	0.273	0.134	0.241	0.159
RW1401	**0.550**	0.547	0.475	0.485	0.443	0.473	0.217	0.281	0.159	0.398	0.203
SCWS1994	**0.691**	0.680	0.624	0.652	0.625	0.643	0.611	0.587	0.422	0.558	0.469
Avg(ρ)	**0.676**	0.648	0.628	0.620	0.604	0.602	0.563	0.543	0.442	0.537	0.365
Dataset	Harmonic score (h) values in all relatedness datasets										
MTurk771	**0.722**	0.692	0.710	0.651	0.657	0.660	0.683	0.594	0.503	0.442	0.510
$MTurk287_{235}$	0.702	0.662	**0.736**	0.718	0.646	0.686	0.634	0.612	0.453	0.418	0.411
WS353Rel	**0.693**	0.609	0.658	0.670	0.625	0.584	0.617	0.563	0.420	0.448	0.315
Rel122	0.581	0.541	**0.591**	0.539	0.517	0.568	0.544	0.465	0.555	0.404	0.420
WS353Full	**0.719**	0.676	0.714	0.718	0.691	0.662	0.646	0.636	0.552	0.401	0.382
MEN	0.776	0.740	**0.800**	0.759	0.739	0.727	0.709	0.681	0.632	0.470	0.436
YP130	0.694	0.682	0.522	0.513	**0.784**	0.543	0.660	0.698	0.606	0.630	0.383
RW2034	**0.520**	0.515	0.445	0.447	0.258	0.445	0.241	0.294	0.137	0.254	0.188
RW1401	**0.534**	0.528	0.470	0.467	0.448	0.460	0.250	0.303	0.162	0.364	0.229
SCWS1994	0.197	**0.198**	0.181	0.182	0.188	0.181	0.192	0.190	0.140	0.148	0.149
Avg (h)	**0.614**	0.584	0.583	0.566	0.555	0.552	0.517	0.504	0.416	0.398	0.342

Some linear combinations of Attract–repel and Paragram-ws models with other methods improve current state-of-the-art results in all datasets by a large margin. This conclusion can be drawn by comparing state-of-the-art results shown in Tables 4 to 7 with their corresponding results shown in Tables A.3 to A.6.

5.11. The new state-of-the-art

We set the new state of the art to answer our RQ3 question as follows.

Attract–repel model sets the new state of the art for the word similarity task (see Tables 4, 5 and 8), being the best overall performing method to tackle this later task. However, Attract–repel model is only a definitive winner method among the family of WE models because it is unable to outperform significantly the most recent OTM and OVM measures (see p-values in Tables 4 and 5).

Paragram-ws model sets the new state of the art in the word relatedness task among the family of WE models, being the best overall performing method to tackle this later task (see Tables 7 and 9). In addition, Paragram-ws model outperforms significantly all OM measures based on WN and the rest of WE models with the only exception of the GloVe model (see p-values in Table 7 and A.6). Thus, Paragram-ws model is a convincing winner among the family of WE models.

WN-RandomWalks is the best performing ontology-based measure in terms of Spearman correlation (see Table 5) for the word similarity task, as well as a convincing winner among the family of OVM methods. WN-RandomWalks outperforms in Spearman correlation all OTM measures in all similarity datasets, in addition to outperform statistically them in all noun relatedness datasets. However, WN-RandomWalks is

unable to outperform significantly all remaining OVM measures and most of best performing OTM methods in the word similarity task.

CoswJ&C (Lastra-Díaz and García-Serrano, 2015b) similarity measure is the best performing OTM measure in Spearman correlation (see Table 5) for the word similarity task, whilst Cai et al. measure is the best one in Pearson correlation (see Table 4). However, neither coswJ&C nor Cai et al. measures are a definitive or convincing winner among the family of OTM methods.

Attract–repel has almost definitively outperformed the family of OM measures in the word similarity task, whilst Paragram-ws model has achieved this later goal for the word relatedness task. Both Attract–repel and Paragram-ws models are able to overcome the strong lexical coverage limitations associated to the family of ontology-based measures based on WN. However, it is interesting to highlight that Attract–repel model partially owes its performance to the use of WN. Thus, WN still being the most used and best performing knowledge base to tackle the problem on word similarity.

Other interesting fact is that most methods which are best suited to estimate the degree of similarity between words perform worst on word relatedness and vice-versa, such as the Attract–repel. However, Paragram-ws model is an exception to this later fact because it is the best performing method for the word relatedness task but it is able to obtain the third best average performance in the word similarity task as shown in Table 8.

Finally, our results show that any linear combination of the best performing similarity and relatedness measures, defined by Attract–repel and Paragram-ws models respectively, significantly improves the performance of its base measure with only two exceptions in the word similarity task. In addition, some linear combinations of Attract–repel and Paragram-ws models with other methods set new state-of-the-art results by a wide margin in all datasets.

5.12. Contradictory results

We obtained several contradictory results in our experiments, confirming the same findings reported in our aforementioned works (Lastra-Díaz and García-Serrano, 2015b,a, 2016), as well as other new ones. For instance, Meng and Gu (2012), Meng et al. (2014) report Pearson correlation values of 0.8804 and 0.8817 respectively with the Seco et al. IC model in the RG65 dataset, whilst we obtained 0.8596 and 0.8486 respectively. Likewise, Cai et al. (2017, table 11) reports Pearson correlation values of 0.91, 0.90. 0.45 and 0.75 in the MC28, RG65, WS-353 and WS-sim (Agirre201) datasets respectively, for the evaluation of their *strategy 2* measure in combination with their intrinsic IC model which is also introduced in Cai et al. (2017). However, we obtained Pearson correlation values of 0.858, 0.872, 0.081 and 0.687 for the evaluation of our software implementations of the Cai et al. (2017, strategy 2) measure and their IC model in the same datasets respectively. These findings confirm again the reproducibility problems in the area. 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, such as done in Appendix B.

6. Conclusions and future work

We have introduced the largest, detailed and reproducible experimental survey on semantic word similarity and relatedness reported in the literature including most of ontology-based semantic similarity measures and word embedding models.

Attract–repel model sets in a statistically significant manner the new state of the art in the word similarity task among the family of WE models, in addition to outperform significantly all ontology-based measures based on WordNet with the only exception of the coswJ&C similarity measure. On the other hand, Paragram-ws model sets in a statistically significant manner the new state of the art in the word relatedness task among the family of WE models with the only exception of the GloVe model, in addition to outperform significantly all ontology-based semantic similarity measures based on WordNet as expected because they are not conceived for this task.

WN-RandomWalks is the best performing OM measure in terms of Spearman correlation for the word similarity task; however, it is unable to outperform significantly all remaining OVM and OTM methods. On the other hand, coswJ&C similarity measure is the best performing OTM measure in terms of Spearman correlation for the word similarity task, whilst Cai et al. (2017) measure is the best one in terms of Pearson correlation. However, neither coswJ&C nor Cai et al. (2017) is a definitive or convincing winner because they are unable to outperform significantly all remaining methods in this aforementioned family.

Attract–repel has almost definitively outperformed the family of ontology-based semantic similarity measures in the word similarity task, whilst Paragram-ws model has achieved it for the word relatedness task. Likewise, both Attract–repel and Paragram-ws models have outperformed significantly all WE models in the word similarity and relatedness tasks, with the only exception of GloVe in this later task. Thus, this work confirms the significant progress achieved by the most recent WE models and the research on ontology-based WE models (WEC) as the mainstream for this line of research.

Our results also confirm that the use of ontologies as WordNet still being the best approach to tackle the word similarity task, as witnessed by the use of WordNet in the most recent and best performing WEC models as Attract–repel. On the other hand, the performance of Paragram-ws model confirms again the distributional approach as the best approach to tackle the word relatedness task. Likewise, our results show that any linear combination of the best performing similarity and relatedness measures, defined by Attract–repel and Paragram-ws models respectively, significantly improves the performance of its combined base measures with only two exceptions in the word similarity task. In addition, some linear combinations of Attract–repel and Paragram-ws models with other methods set new state-of-the-art results in all datasets by a large margin.

Finally, the aforementioned achievements of the state-of-the-art WE models, especially those by Attract–repel, set a breakthrough in this line of research by showing that hybrid corpus-based similarity measures are able to outperform most of well-established ontology-based measures whilst avoid all drawbacks of the last ones, such as their limited lexical coverage, the need of building ontologies, the demand of domain experts, the difficulties in setting universally accepted concepts and relationships, and the difficulties in their upgrading. Thus, we expect that these achievements of the state-of-the-art WE models impact all semantic-aware NLP-based tasks and applications in most knowledge domains by providing semantic similarity measures which would benefit in a complementary way from the availability of curated ontologies and large corpus of non-annotated documents, such as most of text mining applications in the biomedical domain.

As forthcoming activities, we plan to evaluate the performance of state-of-the-art semantic measures in specific semantic-aware applications.

Acknowledgements

We are grateful of Fernando González and Juan Corrales for setting up our UNED Dataverse dataset, Yuanyuan Cai for answering kindly our questions to replicate their IC-based similarity measures and IC models in HESML, and http://clouding.io for their technical support to set up our experimental platform. We also are very thankful to José Camacho-Collados for providing the weighting overlap source code which we have integrated into HESML for measuring the similarity between the NASARI vectors. Finally, we are grateful of the anonymous reviewers for their valuable comments to improve the quality of the paper. This work has been partially supported by the Spanish Ministery of Economy and Competitiveness VEMODALEN project (TIN2015-71785-R), the UPV/EHU (excellence research group) and the Spanish Research Agency LIHLITH project (PCIN-2017-118/AEI) in the framework of EU ERA-Net CHIST-ERA.

Appendix A. Evaluation of averaged models and further p-values

This appendix introduces four tables with the same format that Tables 4–7 which detail the results of the evaluation of the linear combination of all measures with best performing methods in all noun similarity and relatedness datasets. Likewise, this appendix introduces two tables detailing the statistical significance analysis for the evaluation of WE and OVM models shown in Tables 8 and 9. This appendix is provided as supplementary material (see Appendix C).

Appendix B. The reproducible experiments on word similarity

This appendix introduces a detailed experimental setup based on a collection of publicly available software tools (Lastra-Díaz and García Serrano, 2018) and reproducibility resources (Lastra-Díaz et al., 2019), being provided as supplementary material (see Appendix C) with the aim of allowing an exact replication of all our experiments and results, as well as providing our software implementation of all methods evaluated herein.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.engappai.2019.07.010.

References

Adhikari, A., Singh, S., Dutta, A., Dutta, B., 2015. A novel information theoretic approach for finding semantic similarity in wordnet. In: Proc. of IEEE Intl. Technical Conference. IEEE, Macau, China, pp. 1–6.

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

Agirre, E., Cuadros, M., Rigau, G., Soroa, A., 2010. Exploring knowledge bases for similarity. Proc. LREC 373–377.

Agirre, E., Soroa, A., 2009. Personalizing pagerank for word sense disambiguation. In: Proc. of the EACL. ACL, pp. 33–41.

Al-Mubaid, H., Nguyen, H., 2009. Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Trans. Syst. Man Cybern. 39 (4), 389–398.

Aletras, N., Stevenson, M., 2015. A hybrid distributional and knowledge-based model of lexical semantics. In: Proc. of the Fourth Joint Conf. on Lexical and Computational Semantics, pp. 20–29.

Auguste, J., Rey, A., Favre, B., 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In: RepEval@EMNLP. ACL, pp. 21–26.

Avrachenkov, K., Litvak, N., Nemirovsky, D., Osipova, N., 2007. Monte Carlo methods in pagerank computation: When one iteration is sufficient. SIAM J. Numer. Anal. 45 (2), 890–904.

Banerjee, S., Pedersen, T., 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In: Intl. Conf. on Intelligent Text Processing and Computational Linguistics. Springer, pp. 136–145.

Banerjee, S., Pedersen, T., 2003. Extended gloss overlaps as a measure of semantic relatedness. In: Proc. of IJCAI, pp. 805–810.

Banjade, R., Maharjan, N., Niraula, N.B., Rus, V., Gautam, D., 2015. Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. In: Proc. of CICLing, pp. 335–346.

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O., 2007. Open information extraction from the web.. In: Proc. of IJCAI, vol. 7, pp. 2670–2676.

Batet, M., Sánchez, D., 2016. Improving semantic relatedness assessments: Ontologies meet textual corpora. Procedia Comput. Sci. 96, 365–374.

Batet, M., Sánchez, D., 2019. Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. Artificial Intelligence Review 1–19.

Batet, M., Sánchez, D., Valls, A., 2011. An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inform. 44 (1), 118–125.

Ben Aouicha, M., Hadj Taieb, M.A., 2015. G2WS: Gloss-based WordNet and Wiktionary semantic Similarity measure. In: Proc. of ACS/IEEE Intl. Conf. of Computer Systems and Applications, pp. 1–7.

Ben Aouicha, M., Hadj Taieb, M.A., 2016. Computing semantic similarity between biomedical concepts using new information content approach. J. Biomed. Inform. 59, 258–275.

Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A., 2016a. LWCR: multi-layered wikipedia representation for computing word relatedness. Neurocomputing 216, 816–843.

Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A., 2016b. SISR: System for integrating semantic relatedness and similarity measures. Soft Comput. 1–25.

Ben Aouicha, M., Hadj Taieb, M.A., Ben Hamadou, A., 2016c. Taxonomy-based information content and WordNet-wiktionary-wikipedia glosses for semantic relatedness. Appl. Intell. 1–37.

Ben Aouicha, M., Hadj Taieb, M.A., Ezzeddine, M., 2016d. Derivation of "is a" taxonomy from wikipedia category graph. Eng. Appl. Artif. Intell. 50, 265–286.

Ben Aouicha, M., Hadj Taieb, M.A., Ibn Marai, H., 2016e. Wsd-tic: Word sense disambiguation using taxonomic information content. In: Proc. of ICCCI. In: LNCS, vol. 9875, Springer, pp. 131–142.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 3 (Feb), 1137–1155.

Bian, J., Gao, B., Liu, T.-Y., 2014. Knowledge-powered deep learning for word embedding. In: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 132–148.

Blanchard, E., Harzallah, M., Kuntz, P., 2008. A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Proc. of ECAI. IOS Press, pp. 20–24.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2016. Enriching word vectors with subword information. arXiv:1607.04606.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proc. of the ACM SIGMOD, pp. 1247–1250.

Bollegala, D., Alsuhaibani, M., Maehara, T., Kawarabayashi, K.-i., 2016. Joint word representation learning using a corpus and a semantic lexicon. In: Proc. of AAAI, pp. 2690–2696.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30 (1–7), 107–117.

Bruni, E., Tran, N.-K., Baroni, M., 2014. Multimodal distributional semantics. J. Artificial Intelligence Res. 49 (1), 1–47.

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

Cai, Y., Pan, S., Wang, X., Chen, H., Cai, X., Zuo, M., 2018. Measuring distance-based semantic similarity using meronymy and hyponymy relations. Neural Comput. Appl. http://dx.doi.org/10.1007/s00521-018-3766-9.

Cai, Y., Zhang, Q., Lu, W., Che, X., 2017. A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. J. Intell. Inf. Syst. 1–25.

Camacho-Collados, J., Pilehvar, M.T., Navigli, R., 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence 240, 36–64.

Chamberlain, B.P., Clough, J., Deisenroth, M.P., 2017. Neural embeddings of graphs in hyperbolic space. arXiv:1705.10359 [stat.ML].

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.

Chirigati, F., Rampin, R., Shasha, D., Freire, J., 2016. ReproZip: computational reproducibility with ease. In: Proc. of the ACM Intl. Conf. on Management of Data, SIGMOD, vol. 16, pp. 2085–2088.

Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proc. of the 25th Intl. Conf. on Machine Learning. ACM, pp. 160–167.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12 (Aug), 2493–2537.

Cruse, D., 1986. Lexical Semantics. Cambridge University Press, Cambridge, UK.

Davidov, D., Rappoport, A., 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In: Proc. of the 21st Intl. Conf. on Computational Linguistics and the 44th Annual Meeting of the ACL, pp. 297–304.

Deza, M., Deza, E., 2009. Encyclopedia of Distances. Springer.

Dhillon, P.S., Foster, D.P., Ungar, L.H., 2015. Eigenwords: spectral word embeddings.. J. Mach. Learn. Res. 16, 3035–3078.

Dong, L., Srimani, P.K., Wang, J.Z., 2010. WEST: weighted-edge based similarity measurement tools for word semantics. In: Web Intelligence. IEEE Computer Society, pp. 216–223.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A., 2014. Retrofitting word vectors to semantic lexicons. arXiv:1411.4166.

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A., 2015. Retrofitting word vectors to semantic lexicons. In: Proc. of the Conf. of the North American Chapter of the ACL: Human Language Technologies. ACL, pp. 1606–1615.

Faruqui, M., Dyer, C., 2015. Non-distributional word vector representations. arXiv preprint arXiv:1506.05230.

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.

Ganitkevitch, J., Van Durme, B., Callison-Burch, C., 2013 PPDB: The paraphrase database. In: Proc. of HLT-NAACL, pp. 758–764.

Gao, J.B., Zhang, B.W., Chen, X.H., 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. Eng. Appl. Artif. Intell. 39, 80–88.
Garla, V.N., Brandt, C., 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics 13:261.
Georgiev, G.V., Georgiev, D.D., 2018. Enhancing user creativity: Semantic measures for idea generation. Knowl.-Based Syst. 151 (1), 1–15.
Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A., 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In: Proc. of EMNLP, Austin, Texas, pp. 2173–2182.
Glavas, G., Franco-Salvador, M., Ponzetto, S.P., Rosso, P., 2018. A resource-light method for cross-lingual semantic textual similarity. Knowl.-Based Syst. 143, 1–9.
Goikoetxea, J., Agirre, E., Soroa, A., 2016. Single or multiple? Combining word representations independently learned from text and WordNet. In: Proc. of AAAI, pp. 2608–2614.
Goikoetxea, J., Soroa, A., Agirre, E., 2015. Random walks and neural network language models on knowledge bases. In: Proc. of HLT-NAACL, pp. 1434–1439.
Goikoetxea, J., Soroa, A., Agirre, E., 2018. Bilingual embeddings with random walks over multilingual wordnets. Knowl.-Based Syst. 150 (15), 218–230.
Goldberg, Y., Levy, O., 2014. word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv:1402.3722.
Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks. In: Proc. of the 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, pp. 855–864.
Hadj Taieb, M.A., Ben Aouicha, M., Ben Hamadou, A., 2014a. A new semantic relatedness measurement using wordnet features. Knowl. Inf. Syst. 41 (2), 467–497.
Hadj Taieb, M.A., Ben Aouicha, M., Ben Hamadou, A., 2014b. Ontology-based approach for measuring semantic similarity. Eng. Appl. Artif. Intell. 36, 238–261.
Hadj Taieb, M.A., Ben Aouicha, M., Bourouis, Y., 2015. Fm3s: Features-based measure of sentences semantic similarity. In: Proc. of HAIS. In: LNCS, vol. 9121, Springer, pp. 515–529.
Hadj Taieb, M.A., Ben Aouicha, M., Tmar, M., Ben Hamadou, A., 2012. Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring. In: Proc. of ICDKE. In: LNCS, vol. 7696, Springer, pp. 128–140.
Halawi, G., Dror, G., Gabrilovich, E., Koren, Y., 2012. Large-scale learning of word relatedness with constraints. In: Proc. of ACM SIGKDD. ACM, New York, NY, USA, pp. 1406–1414.
Hao, D., Zuo, W., Peng, T., He, F., 2011. An approach for calculating semantic similarity between words using WordNet. In: Proc. of the Intl. Conf. on Digital Manufacturing Automation. IEEE, pp. 177–180.
Harispe, S., Imoussaten, A., Trousset, F., Montmain, J., 2015a. On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies. In: Proc. of the IEEE Intl. Conf. on Fuzzy Systems. Istambul, Turkey, pp. 1–8.
Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 2014. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics 30 (5), 740–742.
Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 2015b. Semantic Similarity from Natural Language and Ontology Analysis. In: Synthesis Lectures on HLT, vol. 8, Morgan & Claypool publishing.
Harris, Z.S., 1954. Distributional structure. Word 10 (2-3), 146–162.
Hassan, S., Mihalcea, R., 2011. Semantic relatedness using salient semantic analysis. In: Proc. of the Twenty-Fifth AAAI Conference. pp. 884–889.
Hearst, M.A., 1992. Automatic acquisition of hyponyms from large text corpora. In: Proc. of COLING, Vol. 2. ACL, pp. 539–545.
Hill, F., Cho, K., Jean, S., Devin, C., Bengio, Y., 2014. Embedding word similarity with neural machine translation. arXiv:1412.6448.
Hill, F., Reichart, R., Korhonen, A., 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. Comput. Linguist. 41 (4), 665–695.
Hirst, G., St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: An Electronic Lexical Database. MIT Press, pp. 305–332.
Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y., 2012. Improving word representations via global context and multiple word prototypes. In: Proc. of the Annual Meeting of the ACL, vol. 1, pp. 873–882.
Jauhar, S.K., Dyer, C., Hovy, E., 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In: Proc. of HLT-NAACL, pp. 683–693.
Ji, X., Ritter, A., Yen, P., 2017. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. J. Biomed. Inform. 69, 33–42.
Jiang, Y., Bai, W., Zhang, X., Hu, J., 2017. Wikipedia-based information content and semantic similarity computation. Inf. Process. Manage. 53 (1), 248–265.
Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy, in: Proc. of Intl. Conf. Research on Computational Linguistics, ROCLING X, pp. 19–33.
Kiela, D., Hill, F., Clark, S., 2015. Specializing word embeddings for similarity or relatedness. In: Proc. of EMNLP, pp. 2044–2048.
Kim, Y., Jernite, Y., Sontag, D., Rush, A.M., 2016. Character-aware neural language models. In: AAAI, pp. 2741–2749.
Lastra-Díaz, J.J., García-Serrano, A., 2015a. A new family of information content models with an experimental survey on WordNet. Knowl.-Based Syst. 89, 509–526.
Lastra-Díaz, J.J., García-Serrano, A., 2015b. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. Eng. Appl. Artif. Intell. 46, 140–153.
Lastra-Díaz, J.J., García-Serrano, A., 2016. A refinement of the Well-Founded Information Content Models with a Very Detailed Experimental Survey on WordNet. Tech. Rep.. UNED, http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement.
Lastra-Díaz, J.J., García Serrano, A., 2018. HESML V1R4 Java Software Library of Ontology-Based Semantic Similarity Measures and Information Content Models. Mendeley Data, v4. http://dx.doi.org/10.17632/t87s78dg78.4.
Lastra-Díaz, J.J., 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, J.J., Goikoetxea, J., Hadj Taieb, M.A., García-Serrano, A., Ben Aouicha, M., Agirre, E., 2019. Word Similarity Benchmarks of Recent Word Embedding Models and Ontology-Based Semantic Similarity Measures. e-cienciaDatos, http://dx.doi.org/10.21950/AQ1CVX.
Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: WordNet: An Electronic Lexical Database. MIT Press, pp. 265–283, ch. 11.
Lesk, M., 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proc. of the Intl. Conf. on Systems Documentation. ACM, pp. 24–26.
Li, Y., Bandar, Z., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. Knowl. Data Eng. 15 (4), 871–882.
Likavec, S., Lombardi, I., Cena, F., 2019. Sigmoid similarity - a new feature-based similarity measure. Information Sciences 481, 203–218.
Lin, D., 1998. An information-theoretic definition of similarity. In: Proc. of ICML, vol. 98, Madison, WI, pp. 296–304.
Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., Hu, Y., 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In: Proc. of the Annual Meeting of the ACL and IJCNLP, vol. 1, pp. 1501–1511.
Liu, Q., Liu, B., Zhang, Y., Kim, D.S., Gao, Z., 2016. Improving opinion aspect extraction using semantic similarity and aspect associations. In: Proc. of AAAI. AAAI Press, pp. 2986–2992.
Liu, X., Zhou, Y., Zheng, R., 2007. Measuring semantic similarity in WordNet. In: Proc. of the 2007 Intl. Conf. on Machine Learning and Cybernetics, vol. 6, IEEE, pp. 3431–3435.
Luong, T., Socher, R., Manning, C.D., 2013. Better word representations with recursive neural networks for morphology. In: Proc. of CoNLL, pp. 104–113.
Manna, S., Mendis, S., 2010. Fuzzy word similarity: A semantic approach using wordnet. In: Proc. of the IEEE Intl. Conf. on Fuzzy Systems. IEEE, Barcelona, Spain, pp. 1–8.
Martinez-Gil, J., 2016. CoTO: A novel approach for fuzzy aggregation of semantic similarity measures. Cogni. Syst. Res. 40, 8–17.
Mazandu, G.K., Chimusa, E.R., Mulder, N.J., 2016. Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Brief. Bioinform. 18 (5), 886–901.
Meng, L., Gu, J., 2012. A New Model for Measuring Word Sense Similarity in WordNet. In: Proc. of the 4th Intl. Conf. on Advanced Communication and Networking, ASTL, vol. 14, pp. 18–23.
Meng, L., Gu, J., Zhou, Z., 2012. A new model of information content based on concept's topology for measuring semantic similarity in WordNet. Int. J. Grid Distributed Comput. 5 (3), 81–93.
Meng, L., Huang, R., Gu, J., 2014. Measuring semantic similarity of word pairs using path and information content. Intl. J. Future Gener. Commun. Netw. 7 (3), 183–194.
Meymandpour, R., Davis, J.G., 2016. A semantic similarity measure for linked data: An information content-based approach. Knowl.-Based Syst. 109, 276–293.
Mihalcea, R., 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In: Proc. of HLT/EMNLP. ACL, pp. 411–418.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. In: Proc. of NIPS, pp. 3111–3119.
Miller, G.A., 1995. WordNet: A lexical database for english. Commun. ACM 38 (11), 39–41.
Miller, G.A., Charles, W.G., 1991. Contextual correlates of semantic similarity. Lang. Cogn. Process. 6 (1), 1–28.
Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P.-H., Vandyke, D., Wen, T.-H., Young, S., 2016. Counter-fitting word vectors to linguistic constraints. In: Proceedings of NAACL-HLT. ACL, San Diego, CA, USA, pp. 142–148.
Mrkšić, N., Vulić, I., Séaghdha, D.Ó., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., Young, S., 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. Trans. ACL 5, 309–324.

Nalisnick, E., Mitra, B., Craswell, N., Caruana, R., 2016. Improving document ranking with dual word embeddings. In: Proc. of the 25th Intl. Conf. Companion on World Wide Web, pp. 83–84.
Navigli, R., Ponzetto, S.P., 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence (ISSN: 0004-3702) 193, 217–250.
Nguyen, K.A., Walde, S.S.i., Vu, N.T., 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. arXiv:1605.07766 [cs.CL].
Nickel, M., Kiela, D., 2017. Poincaré embeddings for learning hierarchical representations. In: Proc. of NIPS, pp. 6341–6350.
Ono, M., Miwa, M., Sasaki, Y., 2015. Word embedding-based antonym detection using thesauri and distributional informationin. In: Proc. of NAACL-HLT, pp. 984–989.
Osborne, D., Narayan, S., Cohen, S.B., 2015. Encoding prior knowledge with eigenword embeddings. Trans. ACL 4, 417–430.
Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proc. of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together, vol. 1501, pp. 1–8.
Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. 40 (3), 288–299.
Pedersen, T., Patwardhan, S., Michelizzi, J., 2004. WordNet::Similarity: Measuring the relatedness of concepts. In: Demonstration Papers At HLT-NAACL 2004. ACL, Stroudsburg, PA, USA, pp. 38–41.
Pekar, V., Staab, S., 2002. Taxonomy learning: Factoring the structure of a taxonomy into a semantic classification decision. In: Proc. of COLING, Vol. 1. ACL, Stroudsburg, PA, USA, pp. 1–7.
Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. Proc. EMNLP 12, 1532–1543.
Perozzi, B., Al-Rfou, R., Skiena, S., 2014. Deepwalk: Online learning of social representations. In: Proc. of the 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. ACM, pp. 701–710.
Petrakis, E., Varelas, G., Hliaoutakis, A., Raftopoulou, P., 2006. X-similarity: computing semantic similarity between concepts from different ontologies. J. Digital Inf. Manag. 4 (4), 233–237.
Pirró, G., 2009. A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. 68 (11), 1289–1308.
Pirró, G., Euzenat, J., 2010. A feature and information theoretic framework for semantic similarity and relatedness. In: Proc. ISWC. In: LNCS, vol. 6496, Springer, Shangai, China, pp. 615–630.
Pirró, G., Seco, N., 2008. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In: On the Move to Meaningful Internet Systems: OTM 2008. In: LNCS, vol. 5332, Springer, pp. 1271–1288.
Quintero, R., Torres-Ruiz, M., Menchaca-Mendez, R., Moreno-Armendariz, M.A., Guzman, G., Moreno-Ibarra, M., 2019. Dis-c: conceptual distance in ontologies, a graph-based approach. Knowledge and information systems 59 (1), 33–65.
Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. 19 (1), 17–30.
Radinsky, K., Agichtein, E., Gabrilovich, E., Markovitch, S., 2011. A word at a time: computing word relatedness using temporal semantic analysis. In: Proc. of the Intl. Conf. on WWW. ACM, pp. 337–346.
Rastogi, P., Van Durme, B., Arora, R., 2015. Multiview LSA: Representation learning via generalized CCA. In: Proc. of HLT-NAACL, pp. 556–566.
Recski, G., Iklódi, E., Pajkossy, K., Kornai, A., 2016. Measuring semantic similarity of words using concept networks. In: Proc. of the 1st Workshop on Representation Learning for NLP, pp. 193–200.
Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of IJCAI, vol. 1, pp. 448–453.
Resnik, P., 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artificial Intelligence Res. 11, 95–130.
Rodríguez, M.A., Egenhofer, M.J., 2003. Determining semantic similarity among entity classes from different ontologies. IEEE Trans. Knowl. Data Eng. 15 (2), 442–456.
Rothe, S., Schütze, H., 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In: Proc. of the Annual Meeting of the ACL and the Intl. Joint Conf. on NLP, pp. 1793–1803.
Rubenstein, H., Goodenough, J.B., 1965. Contextual correlates of synonymy. Commun. ACM 8 (10), 627–633.
Sánchez, D., Batet, M., 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J. Biomed. Inform. 44 (5), 749–759.
Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. Int. J. Semantic Web Inf. Syst. ISWIS 8 (2), 34–50.
Sánchez, D., Batet, M., Isern, D., 2011. Ontology-based information content computation. Knowl.-Based Syst. 24 (2), 297–303.
Sánchez, D., Batet, M., Isern, D., Valls, A., 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl. 39 (9), 7718–7728.
Santos, C.D., Zadrozny, B., 2014. Learning character-level representations for part-of-speech tagging. In: Proc. of ICML, pp. 1818–1826.
Schwartz, R., Reichart, R., Rappoport, A., 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In: Proc. of the Conf. on Computational Natural Language Learning, pp. 258–267.
Sebti, A., Barfroush, A.A., 2008. A new word sense similarity measure in WordNet. In: Proc. of the Intl. Multiconference on Computer Science and Information Technology. IEEE, pp. 369–373.
Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in wordnet. In: Proc. of ECAI, vol. 16, IOS Press, Valencia, Spain, pp. 1089–1094.
Seddiqui, M.H., Aono, M., 2010. Metric of intrinsic information content for measuring semantic similarity in an ontology. In: Proc. of the 7th Asia-Pacific Conf. on Conceptual Modelling, vol. 110, pp. 89–96.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Shima, H., 2011. WS4J Home Page. https://code.google.com/p/ws4j/.
Sinha, R., Mihalcea, R., 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: Intl. Conf. on Semantic Computing, ICSC 2007. IEEE, pp. 363–369.
Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y., 2014. Grounded compositional semantics for finding and describing images with sentences. Trans. ACL 2 (1), 207–218.
Socher, R., Lin, C.C., Manning, C., Ng, A.Y., 2011. Parsing natural scenes and natural language with recursive neural networks. In: Proc. of the 28th Intl. Conf. on Machine Learning, ICML-11, pp. 129–136.
Stanchev, L., 2014. Creating a similarity graph from wordnet. In: Proc. of the 4th Intl. Conf. on Web Intelligence, Mining and Semantics (WIMS'14). Article No. 36. ACM.
Stojanovic, N., Maedche, A., Staab, S., Studer, R., Sure, Y., 2001. SEAL: A framework for developing SEmantic PortALs. In: Proc. of the 1st Intl. Conf. on Knowledge Capture, K-CAP. ACM, pp. 155–162.
Szumlanski, S.R., Gomez, F., Sims, V.K., 2013. A new set of norms for semantic relatedness measures. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'2013), vol. 2, aclweb.org, Sofia, Bulgaria, pp. 890–895.
Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. Line: Large-scale information network embedding. In: Proc. of the 24th Intl. Conf. on World Wide Web, pp. 1067–1077.
Turian, J., Ratinov, L., Bengio, Y., 2010. Word representations: a simple and general method for semi-supervised learning. In: Proc. of the 48th Annual Meeting of the ACL. ACL, pp. 384–394.
Tversky, A., 1977. Features of similarity. Psychol. Rev. 84 (4), 327–352.
Van Miltenburg, E., 2016. WordNet-based similarity metrics for adjectives. In: Proc. of the Global WordNet Conference, pp. 414–418.
Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014a. Knowledge graph and text jointly embeddingin. In: Proc. of EMNLP, pp. 1591–1601.
Wang, Z., Zhang, J., Feng, J., Chen, Z., 2014b. Knowledge graph embedding by translating on hyperplanes. In: Proc. of AAAI, vol. 14, pp. 1112–1119.
Weeds, J., 2003. Measures and Applications of Lexical Distributional Similarity (Ph.D. thesis). Department of Informatics, University of Sussex.
Wieting, J., Bansal, M., Gimpel, K., Livescu, K., 2016. Charagram: Embedding words and sentences via character n-grams. arXiv:1607.02789.
Wieting, J., Bansal, M., Gimpel, K., Livescu, K., Roth, D., 2015. From paraphrase database to compositional paraphrase model and back. Trans. ACL 3, 345–358.
Witten, I.H., Milne, D.N., 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. AAAI Press, pp. 25–30.
Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: Proc. of the Annual Meeting of ACL. ACL, pp. 133–138.
Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., Chen, E., Xu, G., 2017. An efficient wikipedia semantic matching approach to text document classification. Information Sciences 393, 15–28.
Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., Liu, T.-Y., 2014. Rc-net: A general framework for incorporating knowledge into word representations. In: Proc. of CIKM. ACM, pp. 1219–1228.
Yang, D., Powers, D.M., 2006. Verb similarity on the taxonomy of wordnet. In: Proc. of the 3th Intl. WordNet Conf., GWC. Masaryk University, pp. 121–128..
Yu, M., Dredze, M., 2014. Improving lexical embeddings with semantic knowledge. In: Proc. of ACL (Short Papers). ACL, pp. 545–550.
Yuan, Q., Yu, Z., Wang, K., 2013. A new model of information content for measuring the semantic similarity between concepts. In: Proc. of the Intl. Conf. on Cloud Computing and Big Data, CloudCom-Asia 2013. IEEE Computer Society, pp. 141–146.
Zhang, X., Sun, S., Zhang, K., 2018. An information content-based approach for measuring concept semantic similarity in wordnet. Wirel. Pers. Commun. 103 (1), 117–132.
Zhou, Z., Wang, Y., Gu, J., 2008a. A new model of information content for semantic similarity in wordnet. In: Proc .of the Second Intl. Conf. on Future Generation Communication and Networking Symposia, FGCNS'08, vol. 3, IEEE, pp. 85–89.
Zhou, Z., Wang, Y., Gu, J., 2008b. New model of semantic similarity measuring in WordNet. In: Proc. of the 3rd Intl. Conf. on Intelligent System and Knowledge Engineering, vol. 1, IEEE, pp. 256–261.
Zhu, G., Iglesias, C.A., 2017. Sematch: Semantic similarity framework for knowledge graphs. Knowl.-Based Syst. 130, 30–32.
Zhu, G., Iglesias, C.A., 2018. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. Expert Syst. Appl. 101, 8–24.