Lang Resources & Evaluation (2020) 54:645–682
https://doi.org/10.1007/s10579-019-09452-w

ORIGINAL PAPER

Capturing and measuring thematic relatedness

Magdalena Kacmajor[1] · **John D. Kelleher**[2]

Published online: 27 March 2019

Abstract In this paper we explain the difference between two aspects of semantic relatedness: taxonomic and thematic relations. We notice the lack of evaluation tools for measuring thematic relatedness, identify two datasets that can be recommended as thematic benchmarks, and verify them experimentally. In further experiments, we use these datasets to perform a comprehensive analysis of the performance of an extensive sample of computational models of semantic relatedness, classified according to the sources of information they exploit. We report models that are best at each of the two dimensions of semantic relatedness and those that achieve a good balance between the two.

Keywords Semantic relatedness · Thematic relations · Word vector representations · Evaluation datasets

1 Introduction

There are two key dimensions of semantic relatedness. First, concepts can be related because they share many features (consider *mouse* and *rat*), which also implies their membership of same category. Depending on the theoretical perspective, this type of relatedness is known as taxonomic relations or similarity. Second, dissimilar concepts (such as *mouse* and *click*) may be perceived as related due to frequent co-occurrence in some sort of context—for example a temporal, spatial or linguistic one. The resulting relatedness is often referred to as association. The focus of this

✉ Magdalena Kacmajor
magdalena.kacmajor@ie.ibm.com

[1] Innovation Exchange, IBM Ireland, Dublin, Ireland

[2] ADAPT Centre and ICE Research Institute, Technological University Dublin, Dublin, Ireland

paper is on one specific type of associative relationship called thematic relatedness. Thematic relations link concepts playing different, usually complementary roles in the same situation or setting (Lin and Murphy 2001). There is growing evidence from cognitive psychology that thematic relations are crucial to cognitive processes, on a par with taxonomic relations (Jackson et al. 2015).

A taxonomic analysis of a concept is concerned with the inherent features of the concept whereas a thematic perspective deals with the external relations between concepts in a unifying event (Lin and Murphy 2001). Taxonomic relations between concepts are based on a comparison of the concepts' features; concepts that belong to a common taxonomic category share properties and/or functions, and therefore tend to bear physical resemblance. In contrast, thematic relations are formed between concepts performing complementary roles in a common event or theme, which often implies having different (albeit complementary) features and functions.

Many researchers studying the domain of semantic memory (McRae and Boisvert 1998; Jackson et al. 2015) use the term association to refer to non-taxonomic relations and contrast them with taxonomic relations (labelled by these authors as featural/conceptual similarity). However, associative relatedness lacks a precise definition; instead, it is often characterised in terms of free association norms (Nelson et al. 2004), where the strength of an associative link is measured by the probability of one concept evoking another concept. Such an operational definition describes the phenomenon but does not reveal much about the nature of the relation that underlies the frequent co-occurrence (McRae and Boisvert 1998). Words or concepts are often associated because of thematic relatedness, but association may also originate from conventional phrases or idiosyncratic autobiographic memories, or phonological resemblance. For example, when exposed to the cue *coffee*, someone could produce strongly associated response *beard*—because the image of their bearded father drinking coffee is deeply rooted in their memory—whereas for another person, the two concepts would appear completely unrelated.

Moreover, one word may cue another word to come to mind because of taxonomic relations; therefore defining association in terms of free association norms blurs the distinction between associative and taxonomic relatedness. Acknowledging the weakness of the operational definition of association, in this paper we use the following naming convention:

Semantic relatedness	the broadest category that comprises any type of semantic relationship between two concepts.
Taxonomic relations	a subset of relatedness defined as belonging to the same taxonomic category, which involves having common features and functions. In the literature, this type of relatedness is often referred to as similarity.
Non-taxonomic relations	relatedness existing by virtue of co-occurrence of concepts in any sort of context.
Thematic relations	a subset of non-taxonomic relations defined as co-occurrence in events or scenarios, which involves performing *complementary roles*.

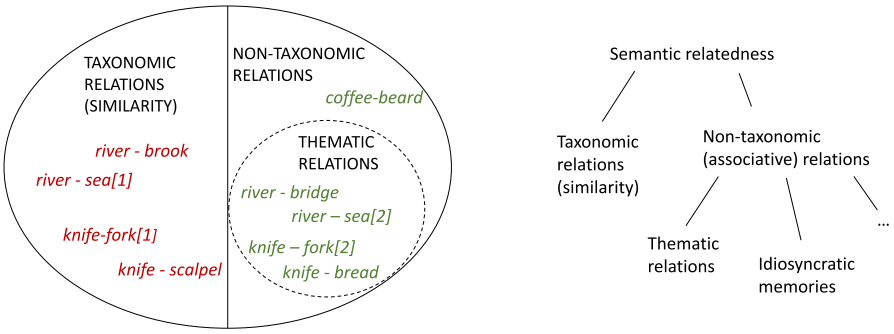

Fig. 1 Subsets of semantic relatedness. The same concept pairs can be linked by two different relatedness types (see pairs marked [1] and [2]). The example of a non-taxonomic and non-thematic relation is an idiosyncratic association produced by one of the authors, whose bearded father is a coffee devotee

Figure 1 shows the subsets of semantic relatedness. It should be noted that, although taxonomic and thematic relations are different and separate types of relatedness, concepts as such may be both taxonomically and thematically related. That is, the distinction applies to the types of relatedness (there is little ambiguity in distinguishing one type of relatedness from the other), but the same pair of concepts can be connected by two different types of relatedness. For example, *doctor* and *nurse* are taxonomically related because they are both members of health professionals category, and also thematically related, because of performing complementary roles, for example during surgery. Which type of relatedness is more salient for a given concept pair, depends on the context, but also on the individual preferences of the observer (Lin and Murphy 2001). In Table 1 we present a few more examples of taxonomic and thematic relations, and explain how two different types of relatedness can co-exist for the same pair of concepts. Some of these examples have been mapped onto the relatedness space in Fig. 1.

The fundamental dichotomy between taxonomic and thematic relatedness has also been recognised in linguistics. For example, in his seminal essay, Jakobson (1956) describes the bipolar structure of language, distinguishing between two types of relations: similarity and contiguity. These relations are explained in terms of two basic operations performed by language users, namely paradigmatic selection and syntagmatic combination. To construct a message, a "communication engineer" selects (substitutable) language units from the common code store, and combines them into higher level contexts. Terms joined by a similarity relation share a substitution set, and thus are subject to selection, whereas members of a contiguity relation (e.g. spatial or temporal) are combined as the constituents of a context. The relations of similarity and contiguity can be applied at different levels of complexity: the constituent parts can be as simple as phonemes within a word context, or as complex as sentences within a context of a broader speech event. Thus the thematic relations we explore in this paper can be perceived as a special case of

Table 1 Examples of concept pairs (C1, C2) that are joined by: a taxonomic relation [1]; a thematic relation [2]; and two relation types [1], [2]

C1	C2	taxonomic [1]	C2	thematic [2]	C2	taxonomic [1], thematic [2]
dog	wolf	Member of same *Canis* genus [1]	bone	Chewed by a dog [2]	cat	Shared categories: carnivore, predator, pet [1]; complementary roles: cat chased by a dog [2]
river	brook	Shared features of a running body of water [1]	bridge	Built over the river [2]	sea	Shared category of a natural water area [1]; complementary roles: river joining the sea [2]
knife	scalpel	Shared features and functions of a cutting blade [1]	bread	A bread knife [2]	fork	Shared features and functions: a piece of cutlery [1]; serving complementary roles during a meal [2]
doctor	surgeon	IS-A relation [1]	ward	Locative relation [2]	nurse	Shared category of health professionals [1]; complementary roles when performing medical duties [2]

Jakobsonian contiguity, and the taxonomic relations as a special case of his similarity, both concerned with words within the context of a phrase or sentence.

To date the majority of computational linguistics research that has been explicit about its definition of semantic relatedness has focused on taxonomic relations (Rada et al. 1989; Budanitsky and Hirst 2006; Hill et al. 2015; Faruqui and Dyer 2015). However, there are good reasons to investigate thematic relationships, including: (a) they are fundamental to cognitive processing, and (b) they are useful for NLP applications.

Estes et al. (2011) provide an extensive review of cognitive research evidencing the critical role of thematic relations in recognizing and understanding words, word pairs, phrases, sentences and whole texts. For example, thematic fit has been shown to constrain the set of words that may occur following a particular word or context (McRae and Matsuki 2009). Given the central role of thematic integration in cognitive processing, including language comprehension, it is worth investigating to what extent thematic relations are captured by existing computational measures of relatedness.

A natural application of thematically-aware models in NLP would be the domain of topic modeling which embraces techniques used to identify topics and estimate their proportion in documents. Thematically-aware models should also be helpful in any tasks requiring word-sense disambiguation, since the correct meaning of a word can be established through identifying its thematic context. Furthermore, the ability to differentiate between taxonomic and thematic relations can lead to enhanced statistical language models. In this last case, both types of relations are important but in a different way: thematic relations express high-probability co-occurrences and thus help to predict the next word, while taxonomic relations indicate which words can be replaced by other words.

The degree of semantic relatedness between two concepts can be expressed as a single number, and the goal of numerous computational measures of semantic relatedness is to produce the best possible estimate of this number. What constitutes the "best" output depends heavily on the type of relatedness that is measured; it is therefore essential that evaluation methods distinguish between taxonomic and thematic relations.

A common way of performing direct intrinsic evaluation of a given semantic relatedness measure is to compare the produced estimates of relatedness to gold standards provided by human judges. Despite certain shortcomings of this approach (comparing *words* rather than *concepts*; ratings performed in isolation from context), gold standard datasets provide valuable quantitative feedback and—at least in theory—allow objective comparison of the performance across very different models of semantic relatedness. Yet, many of commonly used evaluation datasets do not provide direct insight into the nature of the relatedness they measure, yielding a blurred picture of the performance achieved by various computational models.

The limitations of existing evaluation resources have been highlighted by Hill et al. (2015), who have also provided a partial solution to this problem by designing a reliable taxonomic benchmark, SimLex-999[1]. The results from benchmarking on

[1] http://www.cl.cam.ac.uk/~fh295/simlex.html.

Simlex diverge from the results of evaluating against datasets measuring general relatedness, which confirms the claim that targeting specific types of semantic relations has observable impact on evaluation outcomes. However, to date no benchmark has been proposed that would target the other key dimension of semantic relatedness—thematic relations.

In this paper we recommend two datasets that capture thematic relations, yet are not currently used for evaluation purposes. We use these datasets to assess the performance (measured in terms of the correlation with human ratings) of a range of semantic models, including distributional and non-distributional word representations. The results are then analysed from two angles. The goal of Experiment 1, "Learning about datasets from the behaviour of models", is to find evidence that the two candidate thematic datasets measure purely non-taxonomic aspects of relatedness. Our results, together with an analysis of the procedures used when gathering the human assessments of semantic relatedness in each dataset, indicate that the collected ratings predominantly reflect thematic relatedness. The goal of Experiment 2, "Learning about models by evaluating on specialized datasets", is to make practical use of these newly introduced thematic benchmarks and identify best candidate models suitable for specific demands arising from various NLP tasks. Our motivation for the second experiment is fuelled by the assumption that: some applications would benefit from maximizing accuracy in recognizing thematic relations; other tasks will rather require possibly error-free detection of taxonomic relations; yet other applications will need information about both dimensions of relatedness. Therefore, using the thematic datasets side by side with taxonomically-oriented benchmark (Simlex-999), we single out models that (a) are best at capturing thematic relations, (b) are best at capturing taxonomic relations, and (c) achieve best balance between the ability to recognize the two types of semantic relatedness. We find that top performers at one type of relatedness achieve at best mediocre scores at the other dimension; however, exploiting diversified sources of information fosters more balanced systems, and also enhances performance on either benchmark. Finally, we identify a candidate general-purpose benchmark, that is yet another dataset which, according to our evidence, has a good balance between word pairs representing taxonomic and thematic relations.

The paper proceeds as follows: Sect. 2 discusses existing views on representing different aspects of semantic relatedness in NLP. In Sects. 3 and 4 we provide an overview of all datasets and all computational measures of relatedness used in our study. The next three sections focus on experiments, with section 5 outlining data preparation, and Sects. 6 and 7 reporting and analysing the results. Section 8 concludes the paper.

2 Dimensions of semantic relatedness from the NLP viewpoint

Two main computational approaches to modelling semantic relatedness are knowledge-based and distributional. In this section we outline how semantic relatedness and its dimensions are framed in either of these approaches. In the cited literature, terminology used to refer to aspects of relatedness varies from author to

author; for simplicity, in the following discussion we maintain consistency with the naming convention laid out in the Introduction.

2.1 Knowledge-based perspective

Computational lexicons and ontologies, with WordNet (Fellbaum 1998) as the most prominent example, capture taxonomic relations. Consequently, much of the research on knowledge-based measures of relatedness has focused on taxonomic relatedness. Thus, Rada et al. (1989) argue that the hierarchical ("IS-A") links in semantic networks are sufficient to model what they label as conceptual similarity, because such links are defined based on shared features. The focus on hierarchical links among concepts has been followed by multiple authors proposing different variants of WordNet-based similarity measures (Wu and Palmer 1994; Jiang and Conrath 1997; Lin 1998; Leacock and Chodorow 1998). In these approaches, taxonomic relations are defined as the inverse of the length of the path connecting the concepts in WordNet taxonomy, with different ways of normalizing the path.

Resnik (1995) makes a clear distinction between general semantic relatedness and taxonomic relatedness, later reiterated in Budanitsky (1999) and Budanitsky and Hirst (2006): semantic relatedness is defined as the collection of all possible relations between two concepts, and taxonomic relatedness is a subset of these relations limited to "IS-A" links. Other subsets are not clearly distinguished, and the measure designed by Resnik focuses on taxonomic relations only.

An approach proposed by Hirst and St-Onge (1998) represents an attempt to exploit WordNet resources for measuring general semantic relatedness, not restricted to the taxonomic subset. Their lexical chainer, in addition to walking hierarchical links, traverses WordNet paths of meronymy, holonymy and antonymy. However, because of the low proportion of non-taxonomic links in WordNet (see Fig. 2), the ratings of semantic relatedness produced by the lexical chainer are not much different from the results returned from taxonomic similarity measures.

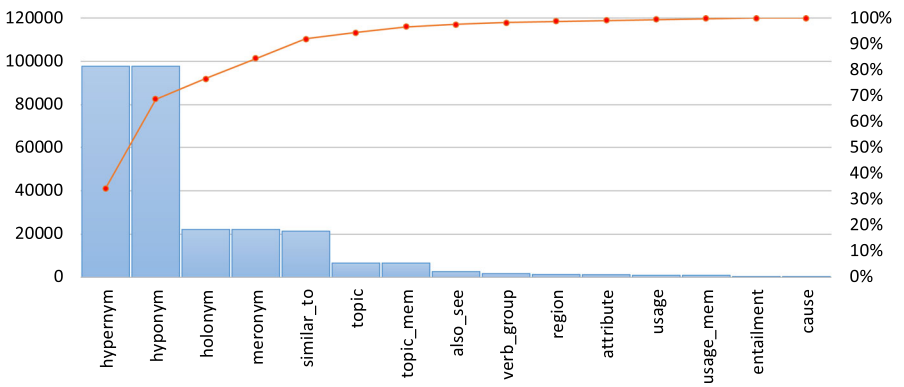

Fig. 2 Counts of semantic pointers in WordNet 3.0. Pareto chart based on data from Finlayson (2015). Bars represent individual counts, and the line represents cumulative total

The problem with both the low number and insufficient range of non-taxonomic links in WordNet has been recognized by Boyd-Graber et al. (2006), who postulate enriching WordNet with non-taxonomic, cross-part-of-speech links. The "radically different" type of information to be captured in these links has been described by the authors in terms of evocation—how strongly one concept brings to mind another concept—and an extensive sample of evocation ratings has been collected from human judges. The evocation dataset is described in more detail in Sect. 3.1.1. Two other wide-scale projects aimed to increase connectivity within WordNet by sense-tagging the glosses (Moldovan and Novischi 2004; Langone et al. 2004)[2]. Words in WordNet glosses have been annotated so as to link the glossed synset with synsets comprising words in the gloss. The resulting gloss relations are cross-part-of-speech and go beyond taxonomic relations.

Morris and Hirst (2004) analyse WordNet limitations in terms of a distinction between classical and non-classical relations, building on Lakoff's concept of classical categories (Lakoff 1987). The notion of classical and non-classical relations can be mapped onto the taxonomic vs. thematic distinction outlined in the Introduction: classical relations are feature-based and well represented in lexical ontologies, while non-classical relations are context dependent and less structural. Furthermore, non-classical relations are not captured in WordNet, yet are obvious for humans and crucial for understanding text.

A number of researchers have noticed the above limitations in WordNet coverage of links among concepts but took a different approach to overcoming the problem: instead of trying to extend WordNet with new relations, they apply the original WordNet-based measures to the Wikipedia hierarchical category graph (Strube and Ponzetto 2006; Zesch and Gurevych 2010). These researchers share the view that taxonomic relations are a subset of general relatedness, and that there is a value in the ability to also recognize non-taxonomic relations that are not encoded in lexical ontologies. Since the Wikipedia category tree does not form a strictly taxonomic hierarchy and the relations among the nodes are not restricted to "IS-A" type, transferring path-based algorithms to Wikipedia should result in modelling broader aspects of relatedness. The models proposed in the cited studies were evaluated against several popular evaluation datasets, but no attempt was made to analyse which aspects of relatedness are captured in these gold standards.

2.2 Distributional perspective

An alternative to knowledge-based measures of semantic relatedness are distributional or corpus-based methods which exploit word co-occurrence statistics derived from large text corpora. Distributional models are also referred to as semantic vector space models, as words extracted from the corpus are represented by a vectors that keep track of the co-occurrences, and their meaning is distributed across multiple dimensions. The proximity of vectors, typically measured by cosine similarity, is interpreted as the relatedness of the words. Vector representations can be constructed by applying factorization to a word-context matrix, or induced using

[2] The Extended Wordnet described in Moldovan and Novischi is available at http://xwn.hlt.utdallas.edu/.

neural language models (Bengio et al. 2003; Mikolov et al. 2013; Pennington et al. 2014). Regardless the implementation or type of contexts used (documents, sliding windows, syntactic contexts), the core assumption is that words that frequently occur in same contexts are taxonomically similar (Harris 1954)—that is, feature overlap can be induced from context overlap. However, empirical evidence shows that the ability of vanilla distributional models (i.e. those based on raw corpus data, not augmented with syntactic or lexical knowledge) to capture taxonomic relations is limited. In the remainder of this section we review several studies that analysed the performance of distributional methods at recognizing different aspects of semantic relatedness.

Agirre et al. (2009) attempted to provide means for evaluating different aspects of semantic relatedness by splitting one of the existing non-discriminative datasets, WordSim-353 (Finkelstein et al. 2001), into two subsets: a taxonomic subset and an non-taxonomic subset. The authors manually classified interword relations from the original dataset using WordNet-style labels. The taxonomic subset contained all the word pairs whose semantic relationship was manually classified as synonymy, antonymy or hypo-/hypernymy. The non-taxonomic subset contained all the word pairs whose semantic relationship was classified as meronymy/holonymy, and also those labelled as "none-of-above", provided that these unidentified relations received human ratings greater than 5 (on a scale from 0 to 10). The need to resort to the "none-of-above" category illustrates the difficulties with the definition of non-taxonomic aspects of relatedness, and confirms the observation that WordNet does not account for many semantic relations that are intuitively obvious for human language users. The same study reported an analysis of the performance of several variants of knowledge-based and distributional models of semantic relatedness, concluding that non-taxonomic relations are best captured by vanilla distributional models, while taxonomic relations are better measured by models using syntactic patterns as the features representing the context.

The finding that using syntactic contexts helps to encode taxonomic relations has been confirmed in a study on neural word embeddings. Levy and Goldberg (2014) propose a modification of Skip-gram model (Mikolov et al. 2013) in which contexts are built from words that are syntactically related to the target word, rather than from all the words surrounding the target word within a given window. The authors conceptualise aspects of semantic relatedness as topical similarity and functional (cohyponymous) similarity, which correspond to thematic and taxonomic relations, respectively. They report their syntactically informed embeddings being less topical (thematic) and more capable of capturing taxonomic relations than the embeddings derived from linear contexts.

Hill et al. (2015) draw a clear distinction between taxonomic and non-taxonomic relations, emphasising the advantages of models that are able to recognize taxonomic relatedness. Hill et al. argue that top-performing distributional models are very accurate when measuring general semantic relatedness but these models are much less capable of recognizing relations defined in terms or shared features or shared category. Hill et al. created a gold standard dataset designed to strictly measure taxonomic relations (Simlex-999), which has quickly become popular among researchers working with word representations. A number of studies have

attempted to enhance the ability of distributional models to capture taxonomic relations, often by incorporating additional sources of knowledge, such as ontologies or syntax. However, while focusing on gains in modelling taxonomic relatedness, little attention has been paid to the accompanying drop in recognizing non-taxonomic relations. The tendency to ignore semantic relations beyond taxonomic ones may have practical reasons: lack of benchmarks targeting non-taxonomic relatedness, and on a more general level, difficulties with defining non-taxonomic semantic relatedness. In this paper, we address the first issue by recommending thematic datasets introduced in Sect. 3.1, and the second issue by drawing attention to the concept of thematic relations that have been extensively studied in cognitive psychology and whose definition can be imported into the domain of natural language processing.

3 Datasets

In our experiments we evaluate a number of computational measures of relatedness against three types of datasets: (1) thematic datasets (evocation dataset and thematic relatedness norms), (2) a dataset known to measure a mixture of taxonomic and thematic relations (USF Free Association Norms) and (3) a dataset targeting taxonomic relations (Simlex-999). The remainder of this section presents all four datasets.

3.1 Thematic datasets

In this subsection we describe two collections of human ratings of thematic relatedness between pairs of concepts. Created using different methodologies, they nevertheless target the same type of relations. Common goals underlying the creation of both datasets include: (1) a focus on non-taxonomic relations, (2) a focus on semantic connections between concepts (as opposed to non-semantic phonological associations between words) and (3) a focus on conventional connotations (representative for a population).

3.1.1 Evocation dataset

The evocation dataset[3] (Boyd-Graber et al. 2006) was created as part of a project that aimed at broadening the range of relation types captured in WordNet (see Sect. 2.1). The dataset was designed to address three main shortcomings of WordNet: (1) the lack of cross-part-of-speech links, (2) the absence of many meaningful relations that do not fit into any of standard ontological labels and (3) the lack of weighted arcs to reflect true semantic distance among related pairs. The three goals underlying the creation of the dataset make it a suitable tool for measuring thematic relations: the cross-POS, syntagmatic links can capture relations between events (verbs) and participants (nouns), or entities (nouns) and their

[3] http://wordnet.cs.princeton.edu/downloads.html.

attributes (adjectives); the quest for new, so far unlabelled relations leads to a shift toward non-taxonomic relations; and adding weights makes the dataset suitable for use as an evaluation tool.

The authors extracted the 1000 most frequent words from the British National Corpus (BNC), preserving the distribution of parts of speech in the lexicon (642 nouns, 207 verbs and 151 adjectives). For each word, they manually selected the most salient or basic synset from WordNet. 120,000 synset pairs were then randomly picked from all the combinations of these synsets, and every pair was annotated by at least 3 judges from a group of 20 undergraduates.

The annotators were asked to rate, on a scale from 1 to 100, how much one synset (or concept) in the pair evokes or brings to mind the other. The instructions stressed that the ratings should reflect the degree of evocation in the general population, not idiosyncratic evocations based on personal history. Furthermore, participants were instructed to only focus on semantic connections and ignore evocations based on phonetic or orthographic resemblance. It was also indicated that evocative relationships do not have to be symmetrical.

Most of the pairs (67%) were rated as unrelated, which is not surprising for random combinations. For the pairs that received at least one non-zero rating, the standard deviation of annotator's ratings per word pair was, on average, 9.25, which is a relatively low value given the scale of ratings (1–100). In our interpretation, the level of agreement among responses indicates that the participants understood the instructions and therefore their ratings reflected non-idiosyncratic semantic relations.

The annotators were not discouraged from assigning high ratings to taxonomically related synset pairs; yet, Boyd-Graber et al. found very poor correlation (0.1 and less) between obtained evocation ratings and selected WordNet-based measures of semantic relatedness. This led them to the conclusion that "evocation is an empirical measure of some aspect of semantic interaction not captured by these similarity methods" (Boyd-Graber et al. 2006).

Since the selected WordNet-based measures rely entirely on hierarchical (taxonomic) links, the results of the analysis conducted by Boyd-Graber et al. indicate that evocation dataset captures primarily non-taxonomic relations. This, in combination with the explicit focus on semantic and non-idiosyncratic aspect of evocation, makes the dataset a promising candidate for a thematic relatedness benchmark. In order to verify the hypothesis about non-taxonomic character of the dataset, we conducted Experiment 1 (see Sect. 6).

3.1.2 Thematic relatedness production norms

The motivation behind Jouravlev and McRae (2015) collecting their datasets was to address the needs of researchers in cognitive psychology interested in the role of thematic thinking in language processing and relatedness judgements. The aim was to identify thematic relations that are conventional, that is salient and well-established in the semantic memory of an average person. This is important because thematic relations, by their very nature, are defined via external events or themes,

and thus are context-dependent and prone to be subjective. The focus on conventional relations is a way to minimize the subjectivity.

Instead of asking people to rate relationship strength between arbitrarily selected concept pairs, Jouravlev and McRae decided to use the production norm method, in which participants are instructed to produce thematically related concepts in response to cue concepts from the provided list. The production frequency of a given response to the same cue concept was used as the measure of the strength of the relation, or the degree of its conventionality.

This methodology is similar to the approach used for collecting free word association norms (Nelson et al. 2004), but the purpose of thematic production norms is different. In case of free word associations, there is no focus or conventionality—idiosyncratic associations are allowed, as well as non-semantic ones (phonetically or orthographically based). In contrast, Jouravlev and McRae used specific instructions to target their thematic production norms at semantic relations that are not based on autobiographical events.

The authors selected 100 concrete concepts commonly used in studies on thematic relations and presented them to 200 students. The participants were given the definition of thematic relations and asked to avoid taxonomically related responses. They were also instructed to respond with nouns only. Several responses for a single cue were allowed.

The final dataset contains cue-responses pairs together with the frequency counts. Only responses produced by at least 10 participants have been considered conventional and included in the dataset, resulting in a collection of 1174 pairs.

3.2 Non-specialised dataset: USF Free Association Norms

USF Free Association Norms database[4] has been collected by researchers at University of South Florida. The dataset consists of over 70,000 cue-response pairs, with the responses produced "under conditions of minimal constraint". The participants were asked to respond with the first word that came to their mind that was meaningfully related to the presented cue. Since they were not restricted to think of any particular type of relatedness, it is reasonable to assume that their unconstrained responses represent a wide spectrum of semantic relatedness.

The values of the association strength assigned to the word pairs are the function of the production frequency. The majority of responses has been normed by a separate group of participants, and thus over 60,000 word pairs have been annotated with both forward (cue to response) and backward (response to cue) association strength. The percentage of nouns, verbs and adjectives is 66%, 17% and 15%, respectively. One in three word pairs (36%) contains words representing different parts of speech.

[4] http://w3.usf.edu/FreeAssociation/.

3.3 Taxonomic dataset: Simlex-999

Simlex-999 (Hill et al. 2015) is a specialised benchmark targeting taxonomic relations. Hill et al. claim that in other evaluation datasets, top-ranked word pairs tend to be both taxonomically and non-taxonomically related, whereas lowest-ranked pairs are not related in any way. Hill et al. argue that therefore it is not possible to identify which aspect of relatedness is captured by computational models that perform well on these other gold standard datasets. In contrast, the word pairs in Simlex-999 represent different types of semantic relatedness, including solely taxonomic (high ranks) and solely non-taxonomic relations (low ranks). Thus, only models that recognize taxonomic relations and ignore non-taxonomic relations perform well on Simlex.

The annotators of Simlex-999 received clear instructions to only assign high ratings to taxonomic relations. The word pairs presented to them originate from USF association norms (Nelson et al. 2004) and have been selected to cover various levels of concretedness, and represent nouns, verbs and adjectives in the proportions consistent with frequencies in the BNC. Given the focus on taxonomic relations, no cross-POS word pairs were included.

4 Computational models

In our first experiment, we use two types of computational models of semantic relatedness: knowledge-based and distributional. In the second experiment we expand the set of investigated models with approaches that go beyond this division by leveraging insights both from statistics and linguistics. This section outlines the three groups of models. For clarity, models based on raw corpus data (Sect. 4.2) are referred to as "vanilla distributional models", to differentiate them from the hybrid approaches. In the remainder of this paper, we will refer to individual models by abbreviated names which are provided in parentheses at the end of each description.

An important factor to consider when defining and working with any model of linguistic semantic relatedness is whether the model (and the dataset it is being evaluated on) uses words or concepts as the basic unit of analysis. The difference is significant, since the same concept may be represented by multiple words (synonymy), and a single word may express multiple meanings (homonymy, polysemy). Knowledge-based approaches that operate directly on a lexicon representation—such as path-based algorithms applied to the WordNet graph—measure semantic relatedness between concepts that were manually distinguished by the linguists who created the lexicon. In contrast, distributional vector representations are automatically derived from co-occurrences of word tokens in text corpora and thus map to unique word forms and not concepts. Knowledge-based vectors described in 4.1.2 also follow the "one vector per word" approach, such that a single vector stores information on all the concepts that might be assigned to the represented word.

Ignoring polysemy and mapping the meaning to words rather than concepts is a known problem of vector space models[5]. On the other hand, referring to words is a convenient simplification that makes it possible to manipulate word representations without engaging in the difficult task of identifying the "right" word sense. With the exception of the evocation dataset (Sect. 3.1.1), datasets described in Sect. 3 are built up by word forms which do not identify concepts. This is suitable for evaluating vector word representations but not knowledge-based approaches that operate on lexical concepts. Since there is typically no information on which concepts were adopted by judges annotating given word pair, a common workaround is to compute the degree of semantic relatedness for all combinations of all senses possible for that word pair, and then select the highest value. We used this workaround when evaluating WordNet-based measures (Sect. 4.1.1).

For the sake of uniformity, we also simplified the evocation dataset, removing all the references to the identified concepts and reducing the dataset to word pairs only (see Sect. 5.1 for the details). This enables evaluation of vector word representations, although at the cost of discarding valuable information (in principle, semantic relatedness occurs between concepts, not words). The level of relatedness between word vectors is measured using cosine similarity, i.e. the cosine of the angle between two word representations positioned in the semantic vector space.

4.1 Knowledge-based methods

The models included in this group rely on human-engineered lexical resources to compute the degree of semantic relatedness between concepts or words. Such resources capture valuable knowledge but are expensive to build. The best-known computational lexical database is WordNet, in which words are organized into synsets (sets of synonyms) that represent distinct concepts. Synsets are interconnected to form a network of semantic (mostly taxonomic) and lexical relations. Other lexical resources propose alternative views on concept's interrelations (Roget's Thesaurus[6], FrameNet[7]) or focus on the linguistic structure of texts (treebanks).

4.1.1 WordNet-based measures

WordNet-based measures (WNM) are the approaches that leverage the graph structure of WordNet in order to measure semantic relatedness of two lexical concepts, as represented by their respective synsets in WordNet. We used modules from WordNet::Similarity API developed by Pedersen et al. (2004) to implement the following selection of seven WNMs:

[5] The complex problem of word sense disambiguation in vector space models, which is out of the scope of this paper, is well explained in Schütze (1998) and Reisinger and Mooney (2010).

[6] http://www.gutenberg.org/ebooks/10681.

[7] https://framenet.icsi.berkeley.edu/fndrupal/.

- Rada et al. (1989) interpret taxonomic relations as the inverse of the path length between two synsets (path).
- Wu and Palmer (1994) scale the path length depending on the position of the "lowest common subsumer" of the compared concepts (wup).
- Leacock and Chodorow (1998) normalize the path length with respect to the maximum depth of the hierarchy (lc).
- Resnik (1995), Jiang and Conrath (1997) and Lin (1998) use frequency statistics derived from a corpus to estimate the probability of encountering an instance of a concept, and thus determine its information content. The degree of relatedness among two concepts is then measured by the amount of information they share (res, jc, lin).
- Hirst and St-Onge (1998) aim to measure semantic relatedness by using a complex semantic distance algorithm to exploit information from both taxonomic and the rare (see Fig. 2) non-taxonomic links existing in WordNet (hso).

4.1.2 Knowledge-based vectors

Knowledge-based (non-distributional) vectors are vector word representations that do not make any use of corpus statistics; instead, the features are extracted from WordNet and other sources of lexical knowledge.

- *Linguistic vectors* (Faruqui and Dyer 2015) are non-distributional vectors constructed from linguistic features that have been derived from multiple knowledge resources, such as WordNet, FrameNet, word-emotion lexicons, Penn Treebank or Roget's Thesaurus. For example, features originating from WordNet are the synsets that a given word belongs to, as well as the related synsets (hypernyms, hyponyms, holonyms etc.). We experiment both with the sparse version downloaded from the authors' repository[8] (vectors of the length of 172,418 dimensions) and with dense versions we obtained by applying SVD (singular value decomposition) to the linguistic matrix to reduce dimensionality, following the methodology described in Faruqui and Dyer (2015) (ling-sparse, ling-svd).

4.2 Vanilla distributional methods

In this group we include word vector representations constructed based on raw corpus statistics, with minimal amount of linguistic preprocessing. Our focus is more on the type of information used to extract the features, not on the particular method of feature extraction; therefore the below list includes both word representations derived through the counting of contexts, and distributed representations learned by neural networks (derived from predicting contexts).

[8] https://github.com/mfaruqui/non-distributional.

Table 2 Training corpora details and dimensionality of vector representations used in experiments

Model	Vector size	Training corpus	Corpus size (# words)
CW	50	Wikipedia	631 M
docNNSE300	300	Clueweb	16 B
ddNNSE300	300	Clueweb	16 B
ddNNSE2500	2500	Clueweb	16 B
glove6B	300	Wikipedia + Gigaword 5	6 B
glove42B	300	Common Crawl	42 B
glove840B	300	Common Crawl	840 B
hpca100	100	Wikipedia + Reuters + WSJ	1.6 B
huang100	50	Wikipedia	1 B
polyen	64	Wikipedia	1.8 B
sg100B	300	Google News	100 B
sg-window5	300	Wikipedia	2 B*
sg-window2	300	Wikipedia	2 B*
sg-deps	300	Wikipedia	2 B*
sp-sparse	9841	Wikipedia + 3 other corpora	5.6 B
sp-500	500	Wikipedia + 3 other corpora	5.6 B
turian100	100	Reuters	37 M

Asterisk denotes estimated size of cleaned Wikipedia dumps from 2014 (the authors of the study do not provide the exact number of words)

Table 2 summarizes training corpora details and the dimensionality of vector representations listed in this and the next section.

- *CW* (Collobert and Weston 2008; Collobert et al. 2011) are vector representations learned using a neural language model which takes as an input a "correct" word sequence *s* observed in the training corpus and a corrupted word sequence *c* generated by replacing one of the words in *s* by a random word, and calculates output scores for both. The objective is to train the word vectors and network combination so that the score returned for each *s* is larger than the score of corrupted word sequence *c*. Throughout the training the weights in each word vector are iteratively adjusted so as to meet this objective. In the experiments we use off-the-shelf vectors[9] trained by Collobert and colleagues using Wikipedia as a corpus (CW).
- *Polyglot*[10] (Al-Rfou et al. 2013) is a variation on CW embeddings, also trained on Wikipedia corpus (polyen).
- *Turian*[11] (Turian et al. 2010) is an implementation of the hierarchical log-bilinear model (Mnih and Hinton 2009)—a probabilistic linear neural model that learns to predict the last word in a context window by linearly combining vector

[9] http://ml.nec-labs.com/senna/.

[10] https://sites.google.com/site/rmyeid/projects/polyglot.

[11] http://metaoptimize.com/projects/wordreprs/.

representations of the preceding words. It uses a hierarchy to filter down the number of performed computations for optimization purposes (turian100).

- Huang et al. (2012) use a combination of window contexts and a document context, training two neural networks against a joint training objective. The input to the first network is a context window that scans through the corpus; the input to the second vector is a weighted average of all the vectors in the document. The training objective is as in Collobert et al. (2011), with the score comprising the outputs of both networks. We trained vectors used in our experiments on Wikipedia corpus, using software shared by the authors[12] (huang100).
- *Document based NNSE*[13] (Murphy et al. 2012) vectors were obtained by applying matrix factorization to a matrix constructed from document co-occurrence counts. The authors use matrix decomposition algorithm called Non-Negative Sparse Embedding (NNSE) method, which is a variation on Non-Negative Sparse Coding (Hoyer 2002) and returns a sparse embedding for each word (that is, for each row in the input matrix) (docNNSE300).
- *SkipGram* (Mikolov et al. 2013) is an efficient neural network language model in which the hidden non-linear layer is removed, simplifying the architecture and reducing computational complexity. The training objective is to predict words within a specified window around the input word. Vectors downloaded from the author's website[14], were trained on 100 billion words of Google News dataset (*sg100B*). In the experiments we also use SkipGram vectors trained by Levy and Goldberg (Levy and Goldberg 2014) on Wikipedia corpus to provide comparison to their dependcy-based SkipGram version (described in Sect. 4.3), and to investigate the effect of window size (*sg-window2*, *sg-window5*).
- *HPCA* (Lebret and Collobert 2014) are word vector representations learned via Hellinger PCA transformation of word co-occurrence matrix obtained through simply counting words over a corpus (*hpca100*).
- *GloVe* (Pennington et al. 2014) model combines count-based and prediction-based approaches. It constructs the matrix of ratios of co-occurrence probabilities, and trains a neural language model on these ratios. Thus, the model directly encodes global corpus statistics. We use off-shelf embeddings trained on corpora of different sizes (glove840B, glove40B, glove6B).

4.3 Other approaches

The third group is comprised of models that combine distributional and knowledge-based approaches or incorporate alternative sources of information. Enriching distributional models with syntactic or ontological information is motivated by more

[12] http://ai.stanford.edu/∼huang/.

[13] http://www.cs.cmu.edu/∼bmurphy/NNSE/.

[14] https://code.google.com/archive/p/word2vec.

or less explicitly stated aspiration to enhance their ability to capture taxonomic relations.

- *Dependency-based distributional vectors* rely on corpus statistics but in contrast to models listed in Sect. 4.2, they require information about the linguistic structure of the corpora text. Co-occurrence is defined in terms of dependency relations between words, not in terms of a linear window. Syntactic contexts can be employed both in count-based and in prediction-based approaches. In our experiments we use dependency-based NNSE—a linguistically informed version of NNSE model (Murphy et al. 2012) in which the input matrix is constructed from dependency counts—and dependency-based SkipGram[15] (Levy and Goldberg 2014) in which neural network language model is trained using syntactic contexts (ddNNSE300, ddNNSE2500, sg-deps).
- *Symmetric Patterns* (Schwartz et al. 2015) are sequences of words and wildcards (such as "X *and* Y", "X *of the* Y"), and vector representations of words have been derived from the co-occurrence of these lexico-syntactic patterns. Both sparse and dense versions have been made available by the authors[16] (sp-sparse, sp-500).
- *RWSGwn*[17] (Goikoetxea et al. 2015) are word embeddings obtained by applying neural network algorithm to a pseudo-corpus generated through random walks over a WordNet graph. The random walk algorithm is based on PageRank, and the produced pseudo-sentences are fed into SkipGram model to induce vector representations of words. Importantly, the graph used for generating the pseudo-corpus is derived from WordNet with gloss relations, which means that final embeddings encode far richer knowledge than information captured in taxonomic links in WordNet (RWSGwn).
- *Concatenated vectors* are constructed as simple concatenation of vector representations trained using complementary approaches. Faruqui and Dyer (2015) append their linguistic vectors to SkipGram vectors, reporting improved performance across several evaluation datasets. In our experiments we test novel combinations of the best performing distributional models (SkipGram, GloVe and NNSE) with linguistic vectors, Symmetric Patterns vectors and RWSGwn.

5 Datasets preparation

Three of the four datasets described in Sects. 3.1 and 3.2 have not been specifically designed for the purpose of evaluating the performance of computational measures of relatedness, hence some degree of data preprocessing was required. In this section we describe steps taken to adapt them as evaluation tools.

[15] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/.

[16] http://homes.cs.washington.edu/~roysch/papers/sp_embeddings/sp_embeddings.html.

[17] http://ixa2.si.ehu.eus/ukb/.

5.1 Preparing the evocation dataset

In the evocation project, annotators rated relatedness between concepts (synsets) rather than words: each initial word selected from BNC was presented to them together with the list of words from the corresponding synset. The original dataset provides the annotated pairs in two formats: as sense-key pairs and as word-sense pairs (see two first columns in Table 3). Neither of these formats is suitable for evaluating distributional word representations which map to word forms in text corpora rather than to concepts; therefore in our evaluation dataset we only consider the initial words. Table 3 shows examples of the annotated pairs as provided by Boyd-Graber et al. (2006) (supercolumns 1 and 2) and after adapting to the evaluation dataset used in our experiments (supercolumn 3).

From the total 119,668 word pairs we selected the 38,735 pairs annotated by at least 5 judges, taking the average of the raw ratings assigned by each annotator. From the resulting set we dropped 25,529 pairs (66%) that have been assessed as unrelated, as well as 30 pairs whose members were instances of phrasal verbs written as a WordNet-style collocation (words joined with an underscore). The remaining 13,176 word pairs are used in our experiments.

5.2 Preparing the thematic relatedness norms dataset

Thematic relatedness production norms consist of cue-responses pairs together with the counts indicating how often each concept was returned as the first, second or third response. Synonymous responses to the same cue have been merged in the original dataset (for example, *flight attendant* and *stewardess* returned as the responses to the cue *airplane*), and their counts have been added. In such cases, we take the most frequent word representing the synonymous response and discard the remaining ones. If the most frequent response is a compound term, we take the second most frequent word from the synonymous group. If there is no single-word response for a cue, we discard the whole entry. The final dataset used in our study comprises 1,122 word pairs.

For each cue-response pair, the final frequency count is weighted by the order of producing the response. Counts of words produced first are multiplied by 3, and counts of words given as a second response are multiplied by 2.

5.3 Preparing the USF Free Association dataset

The values of forward and backward association strength provided for word pairs in USF Free Association database are different. For the purpose of the comparison with the ratings produced by computational measures presented in Sect. 4, which assume symmetrical relatedness between words, we take the average of the forward and backward strength. Thus, out of the total 72,176 cue-response pairs in the database, we discard 8,557 pairs for which only forward association strength has been provided. From the remaining set, we remove 2,005 pairs that represent parts of speech other than nouns, verbs or adjectives. The final dataset consists of 61,526 word pairs.

Table 3 Example entries: formats used in the original evocation dataset (Boyd-Graber et al. 2006) (supercolumns 1 and 2) and in our evaluation dataset (supercolumn 3)

(1) Sense-key pair		(2) Word-sense pair		(3) Word pair	
Old%3:00:02::	Speech%1:10:05::	Old.a.01	Manner_of_speaking.n.01	Old	Speech
Possible%3:00:04::	Still%3:00:01:nonmoving:00	Potential.a.01	Inactive.s.10	Possible	Still

6 Experiment 1: learning about datasets from the behaviour of models

In our first experiment we verify whether the two candidate thematic benchmarks described in Sect. 3.1 indeed capture non-taxonomic relatedness. The idea is to use the performance patterns of different groups of computational models as the indicators of the type of relatedness measured by a given dataset. The performance is defined as the agreement between computed degree of relatedness and human judgment, expressed with Spearman rank correlation coefficient.

We use two groups of models of semantic relatedness presented in Sect. 4: knowledge-based and vanilla distributional. As discussed in Sect. 2, models from the first group (knowledge-based) use information embedded in taxonomic links of lexical ontologies, and therefore are able to recognize taxonomic relations but not non-taxonomic relations, such as thematic relations. Models from the second group derive semantic distance from word co-occurrences in large corpora and, as shown empirically, capture broader relatedness (Agirre et al. 2009; Hill et al. 2015; Levy and Goldberg 2014).[18]

Knowing the capabilities and limitations of each group of computational models, we expect that if an evaluation dataset assigns high ratings only to non-taxonomic relations, then knowledge-based measures should score lower on it than distributional models. A lack of difference between average performance of the two groups of models would suggest that high ratings in the evaluation dataset are assigned to both taxonomically and non-taxonomically related pairs of words.

We analyse the difference in performance scores obtained by knowledge-based and vanilla distributional models against the two candidate thematic benchmarks: evocation dataset (*evoc*) and thematic relatedness production norms (*themrel*). To gain a more comprehensive picture, we perform similar tests using datasets representing other profiles of semantic relatedness: a non-specialized dataset derived from USF Free Association database that presumably captures all types of relations (*usf*), and the taxonomically oriented Simlex-999 (*simlex*).

6.1 Procedure and results

In order to run the experiment using full versions of the four evaluation datasets, it was necessary to restrict the set of evaluated knowledge-based models to the linguistic vectors (ling-sparse, ling-svd) and Wordnet hso model, because other WordNet based measures can only be applied to subsets of word pairs in these datasets (this issue is discussed in more detail in Sect. 6.2).

To obtain estimates of the degree of relatedness from vector models, we computed cosine similarity between vectors representing each word pair. For the WordNet-based measure hso, the degree of relatedness was calculated according to the algorithm proposed in Hirst and St-Onge (1998). Thus, for each dataset, each model yielded a list of relatedness ratings. Next, for every model we determined

[18] At this point, we do not use models that exploit additional or combined sources of information, because the profile of relatedness encapsulated in these models is more obscure and their value as indicators of a dataset's profile is limited. Their capabilities will be investigated in Sect. 7.

Spearman correlation between the computed estimates and human ratings in each of the four datasets.

The resulting performance scores are presented in Table 4. Missing words are ignored in correlation calculation, that is, each model is evaluated against the word pairs covered by its vocabulary (average percentage of missing words per dataset is provided in the table). Since correlation values received for different datasets fall into different scales, we perform range normalization of results achieved by all the models on a given dataset, such that the performance scores are scaled in the range [0,1]. Our further analysis is based on the normalized results, presented on the right side of the table.

For each dataset, we applied independent samples t-test to compare mean performance scores obtained by the samples of distributional and knowledge-based models. F-tests had been run beforehand to determine whether the variances between the two compared samples are equal or not. Based on the results of F-tests, *usf* dataset was tested with Welsch's adaptation of t-test assuming unequal variances, and other datasets—with Student's t-test assuming equal variances and unequal sample size.

Our null hypothesis assumes no difference between mean scores of distributional and knowledge-based samples (which is the likely outcome when the evaluation dataset assigns high ratings for both taxonomically and non-taxonomically related word-pairs). The alternative hypothesis (H_{alt}) for *themrel* and *evoc* claims a difference in favour of distributional models—the pattern expected for thematically oriented benchmarks. H_{alt} for *simlex* predicts a difference in the opposite direction, i.e. better performance of knowledge-based methods. Finally, H_{alt} for *usf* is nondirectional, since we do not have any information suggesting which sample (if any) might exhibit higher mean scores.

Table 5 summarizes the assumptions and results of the test. With significance level set to 0.05, the null hypothesis has been rejected for the two thematic datasets (*themrel* and *evoc*) and the taxonomically oriented *simlex*. As regards *usf* dataset, mean score difference was not sufficient for rejecting null hypothesis. Figure 3 shows average performance within each of the two groups of computational measures of relatedness, illustrating the contrasting behaviour of models and its dependence on the dataset type they are evaluated against.

6.2 Discussion

The results of the experiment are consistent with the hypothesis that the two thematic datasets assign high relatedness ratings only to non-taxonomic links. Knowledge-based methods are unable to detect these links, which explains extremely low correlation between relatedness ratings returned by theses measures and gold standard ratings provided in *themrel* and *evoc*. Distributional methods recognize both taxonomic and non-taxonomic relatedness, and therefore their performance on these datasets is relatively much better (although in absolute terms, their scores on thematic datasets are still not high).

Table 4 Performance scores of vanilla distributional and knowledge-based models evaluated against two thematic datasets (*themrel* and *evoc*), non-specialized dataset (*usf*) and taxonomic dataset (*simlex*)

	Raw results				Normalized results			
Dataset	Themrel	Evoc	Usf	Simlex	Themrel	Evoc	Usf	Simlex
Words total	1,122	13,176	61,526	999	1,122	13,176	61,526	999
Missing words	2.6%	0.8%	3.1%	0.8%	2.6%	0.8%	3.1%	0.8%
Distributional:								
CW	0.16	0.10	0.25	0.27	0.51	0.44	0.50	0.26
turian100	0.07	0.05	0.12	0.22	0.10	0.18	0.00	0.14
polyen	0.11	0.08	0.22	0.24	0.28	0.33	0.41	0.19
hpca100	0.10	0.00	0.18	0.16	0.22	0.00	0.22	0.00
huang100	0.19	0.15	0.25	0.28	0.61	0.66	0.51	0.28
glove6B	0.24	0.22	0.34	0.37	0.88	0.94	0.88	0.50
glove42B	0.27	0.18	0.34	0.37	1.00	0.79	0.89	0.51
glove840B	0.25	0.23	0.37	0.41	0.92	1.00	0.98	0.59
sg100B	0.22	0.19	0.37	0.44	0.79	0.80	1.00	0.67
docNNSE300	0.19	0.15	0.25	0.27	0.61	0.62	0.52	0.26
sg-window5	0.16	0.21	0.33	0.37	0.48	0.88	0.83	0.49
sg-window2	0.14	0.16	0.32	0.41	0.41	0.69	0.79	0.61
Knowledge based:								
ling-svd	0.08	0.06	0.24	0.58	0.12	0.23	0.46	1.00
ling-sparse	0.05	0.01	0.25	0.57	0.00	0.04	0.50	0.97
hso	0.10	0.04	0.24	0.45	0.22	0.14	0.45	0.70

"Words Total" is the total number of word pairs in each dataset. "Missing Words" is the average percentage of missing words per dataset

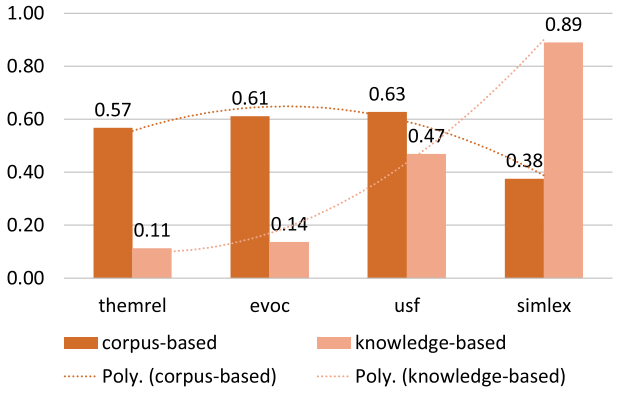

Fig. 3 Average performance scores obtained by vanilla distributional and knowledge-based models on thematic relations norms, evocation, free association norms and Simlex-999. Polynomial trendlines (*Poly.*) are added to accentuate performance patterns

Table 5 T-test results of comparing mean scores obtained by our sample of distributional models (*d*) and our sample of knowledge-based models (*kb*) against four evaluation datasets

Dataset	Themrel*	Evoc*	Usf**	Simlex*
H_{alt}	$\overline{d} > \overline{kb}$	$\overline{d} > \overline{kb}$	$\overline{d} \neq \overline{kb}$	$\overline{d} < \overline{kb}$
t-statistics	2.610	2.540	1.703	3.848
p-value	0.011	0.012	0.117	0.001
H_0 rejected?	Yes	Yes	No	Yes

Null hypothesis assumes no difference between sample means

*One-tailed Student's t-test assuming equal variances, unequal sample size

**Two-tailed Welch's t-test assuming unequal variances, unequal sample size

When evaluated against *usf* dataset, which presumably allocates high ratings to both thematic and taxonomic relations, knowledge-based measures achieve considerably better scores (comparable to the performance of distributional methods), as the taxonomic portion of associations captured in the USF database is visible to them. Finally, the ability of vanilla distributional models to capture thematic relations becomes a disadvantage when evaluating on *simlex*, which has been intentionally designed to penalize models that assign high ratings to non-taxonomic relations (Hill et al. 2015). As a consequence, knowledge-based measures outperform distributional models on this dataset.

In another version of this experiment, we broaden the range of knowledge-based methods to include classical graph-based WordNet-based measures (see 4.1.1) which travel along hierarchical links in the ontology. Since hierarchical links only exists for words that share the same part of speech, and moreover, there is no hierarchical structure for adjectives in WordNet, these methods can only supply ratings for noun-noun and verb-verb pairs. In order to include WNM in our comparison, we re-ran the evaluation of all the models against subsets of the datasets, that is excluding word pairs that are not connected in the WordNet graph. We found that including graph-based measures did not change the outcome of the t-test. Interested readers may find full results of evaluating models on pruned versions of datasets in "Appendix A".

To sum up, our experiment rested on the assumption that knowledge-based models are capable of capturing taxonomic relations, while distributional methods capture general semantic relatedness. This assumption is motivated theoretically given the sources of information utilized in each of these approaches, and also supported by empirical evidence (Agirre et al. 2009; Hill et al. 2015; Levy and Goldberg 2014). We harnessed this knowledge to demonstrate that the thematic relatedness production norms (*themrel*) and evocation dataset (*evoc*) could be useful as specialised benchmarks that selectively capture non-taxonomic relations.

7 Experiment 2: learning about models by evaluating on specialized datasets

In Experiment 1 we tested the type of semantic relatedness captured in the thematic datasets. This test was possible because each of the models selected for the study could be unambiguously classified as either knowledge-based or distributional. Based on theoretical and empirical evidence coming from past research, we consider the relatedness profile of these models as known, which allows us to draw conclusions about the relatedness encapsulated in evaluation datasets.

However, the range of computational approaches to modelling relatedness is by no means limited to pure knowledge-bases and distributional methods. Many approaches use both sources of information (e.g. concatenated vectors) or reach for additional sources, such as dependency relations (Murphy et al. 2012; Levy and Goldberg 2014), symmetric patterns (Schwartz et al. 2015) and gloss relations (Goikoetxea et al. 2015). Based on the evidence coming from Experiment 1, as well as the description of the procedure of obtaining human ratings of relatedness, we now consider dataset's relatedness profile as known, and conduct an experiment to draw conclusions about the relatedness captured by models whose classification is not obvious.

Concretely, we use the thematic datasets, along with Simlex-999 as taxonomic benchmark, to evaluate three aspects of model performance: (a) recognizing thematic relations; (b) recognizing taxonomic relations; and (c) finding a happy medium between the ability to detect the two types of relatedness. Each of these evaluations can be useful for researchers concerned with various applications of models of semantic relatedness. For example, those studying topic modelling would be interested in methods that perform best on thematic benchmarks while those focused on dictionary generation would care mostly about models' capability to recognize taxonomic relations. The third evaluation—identifying the model coping with both types of relatedness—would be valuable for tasks where both taxonomic and thematic relations matter, and the researchers cannot afford to neglect either aspect. This evaluation criterion is relevant in multitask learning (Collobert and Weston 2008), but should be also useful in statistical language modelling. According to Jakobson (Jakobson 1956), the processes of paradigmatic selection and syntagmatic combination are both continually active and intertwining in normal verbal behaviour. The first process requires understanding the internal, structural relations of similarity, and the second process requires understanding the external, operational relations of contiguity (see Sect. 1); failing to grasp either pole results in abnormal speech. The results provided in Table 7 suggest that maximizing accuracy at one dimension comes at the cost of deteriorated performance at the other dimension; since it may be impossible to excel at both, we attempt to find a compromise by identifying best balanced models. We anticipate that acquiring competencies that are crucial for human language users would be beneficial for computational language models. Thus, in this paper we propose using harmonic mean for assessing the level of balance in capturing the two aspects of relatedness.

7.1 Procedure and results

In Experiment 2 we included all types of vector representations listed in Sect. 4. In order to produce concatenated vectors, we built on the approach taken by Faruqui and Dyer (2015) who appended their linguistic vectors to SkipGram embeddings and observed improved performance on several evaluation tasks. Specifically, for this experiment we selected the three vanilla distributional word representations that performed best in Experiment 1 (SkipGram, GloVe and NNSE) and tested their combinations with vectors that use information beyond raw corpus data: linguistic vectors, symmetric pattern vectors[19] and RWSGwn. As for non-vectorial, graph-based measures of semantic relatedness, we included hso (Hirst and St-Onge 1998) which provides relatedness ratings for cross-POS word pairs and can be evaluated using full versions of the datasets.

All the models were evaluated against *themrel*, *evoc* and *simlex*. To identify models that find best balance in recognizing taxonomic and thematic relations, we apply weighted harmonic mean to range-normalized results (as explained in the previous section, range normalization is necessary when comparing model performance across datasets, because the range of results obtained on thematic datasets is not compatible with the scale of scores obtained on *simlex*). Each of the two thematically oriented datasets is assigned the weight of 1, and *simlex* is assigned the weight of 2:

$$weighted\ harmonic\ mean = \frac{1+1+2}{\frac{1}{themrel\ score} + \frac{1}{evoc\ score} + \frac{2}{simlex\ score}} \quad (1)$$

Table 6 shows the results ordered from best to worst with regard to the harmonic mean. In order to provide the full picture of best achievers across different datasets, we present a complete list of evaluated models.

7.2 Discussion

According to the ranking by the harmonic mean (Table 6), the best equilibrium in recognizing thematic and taxonomic relations is achieved by concatenating RWSGwn and GloVe vectors, which yields a combination of three sources of information: raw corpus data, lexical ontology, and gloss relations. More generally, the highest performance in terms of the harmonic mean is observed for concatenated vectors (marked in the table in italics), which dominate the block of top 13 models.

Combining different sources of information not only contributes to more balanced systems but also leads to improved ability to capture either aspect of relatedness. Table 7 shows performance ranks of all the models with respect to each of the three specialised datasets. With few exceptions, the top three performers on each dataset are concatenated vectors.

[19] In an analogous attempt, Schwartz et al. (2015) compute linear combination of relatedness ratings returned by their symmetric pattern model and Skipgram, and report improved performance on Simlex-999. However we find that simple concatenation of vectors yields better results (correlation of 0.58 vs. 0.56).

Table 6 Spearman correlations between relatedness ratings returned by computational models and human ratings collected in thematic relatedness norms (*themrel*), evocation (*evoc*) and Simlex-999 (*simlex*) datasets

No.	Model	Type	*Har-mean*	*Themrel*	*Evoc*	*Simlex*
1	*RWSGwn+glo*	*a*	*0.87*	***0.26*** $^{(2)}$	***0.26*** $^{(1)}$	*0.50*
2	*sp-sparse+sg*	*a*	*0.80*	*0.20*	*0.17*	*0.58*
3	*RWSGwn+sg*	*a*	*0.79*	*0.21*	*0.19*	*0.53*
4	*RWSGwn+nnse*	*a*	*0.79*	*0.20*	*0.19*	*0.53*
5	RWSGwn	rw	0.76	0.19	0.19	0.52
6	*sp-sparse+glo*	*a*	*0.73*	***0.26*** $^{(3)}$	***0.23*** $^{(2)}$	*0.42*
7	*ling-svds+sg*	*a*	*0.72*	*0.23*	*0.19*	*0.45*
8	*ling-svds+glo*	*a*	*0.71*	*0.25*	***0.23*** $^{(3)}$	*0.41*
9	glove840B	d	0.71	0.25	0.23	0.41
10	sg100B	d	0.70	0.22	0.19	0.44
11	*sp-sparse+nnse*	*a*	*0.69*	*0.20*	*0.11*	***0.59*** $^{(1)}$
12	*ling-sparse+glo*	*a*	*0.67*	*0.15*	*0.15*	***0.59*** $^{(2)}$
13	*ling-svds+nnse*	*a*	*0.64*	*0.21*	*0.11*	*0.50*
14	glove6B	d	0.62	0.24	0.22	0.37
15	glove42B	d	0.62	**0.27** (**1**)	0.18	0.37
16	ddNNSE2500	d	0.59	0.21	0.11	0.46
17	sg-window2	d	0.54	0.14	0.16	0.41
18	sg-window5	d	0.54	0.16	0.21	0.37
19	sp-sparse	l	0.53	0.13	0.10	0.54
20	sg-deps	l	0.45	0.12	0.10	0.45
21	huang100	d	0.37	0.19	0.15	0.28
22	docNNSE300	d	0.36	0.19	0.15	0.27
23	ddNNSE300	d	0.33	0.11	0.06	0.37
24	CW	d	0.32	0.16	0.10	0.27
25	ling-svds	kb	0.28	0.08	0.06	**0.58** (**3**)
26	hso	kb	0.27	0.10	0.04	0.45
27	polyen	d	0.23	0.11	0.08	0.24
28	turian100	d	0.14	0.07	0.05	0.22
29	depNNSE300	l	0.13	0.07	0.02	0.40
30	*ling-sparse+sg*	*a*	*0.10*	*0.06*	*0.02*	*0.57*
31	ling-sparse	kb	0.06	0.05	0.01	0.57
32	*ling-sparse+nnse*	*a*	*0.00*	*0.05*	*0.02*	*0.57*

Table 6 continued

No.	Model	Type	*Har-mean*	*Themrel*	*Evoc*	*Simlex*
33	hpca100	d	0.00	0.10	0.00	0.16

Weighted harmonic mean (*har-mean*) has been taken after scaling the results in the range [0,1]. Italic highlights blocks of concatenated vectors. First, second and third best result obtained on each dataset is marked with a superscript with the respective number

Type symbols:

kb = knowledge-based models

d = vanilla distributional models

l = linguistically informed distributional vectors (dependency-based and symmetric patterns)

a = concatenated vectors

rw = RWSGwn

It is, however, apparent that it is hard to reconcile the ability to capture taxonomic relations with the ability to recognize thematic relations. The concatenated model that scores best on *themrel* and second best on *evoc* (No. 1 in Table 6) is placed at position 13 on *simlex*. The combination that performs best on *simlex* (No. 11) is ranked as 12th on *themrel* and 19th on *evoc*.

The lack of balance is exhibited most strongly by models that obtain high scores on *simlex*. Comparing the rank by harmonic mean with the ranks by performance on specialised datasets, we find that the mean values are highly correlated with the scores obtained on thematically-oriented datasets (Spearman's rho of 0.88 and 0.83), and poorly correlated with the scores obtained on taxonomically-oriented Simlex-999 (Spearman's rho of 0.3).[20] This means that best achievers on simlex tend to be penalized more heavily by the harmonic mean which is the type of average that unfavours large differences between its arguments (Kelleher et al. 2015).

This is consistent with the notion that knowledge-based models selectively capture taxonomic relations, and distributional approaches encode general relatedness and thus are not limited to a single aspect. In other words, it is possible to find models that are excellent at modelling taxonomic relatedness while being almost completely blind to thematic relations, but we have not identified models that selectively detect thematic relatedness without recognizing taxonomic links.

Furthermore, ranks presented in Table 7 reveal that the impact of training corpora size on the performance of distributional word representations is important but not ultimately deciding. Sizes of corpora used for training the models under evaluation vary significantly (see Table 2). In general, vectors trained using massive amounts of data score better than those using small or medium corpora, at least in terms of the harmonic mean. However, Glove trained on 840 billion words (glove840B) is outperformed on *simlex* by Skipgram vectors induced on 100 billion words (sg100B) or even just 2 billion (sg-window2). On the other hand, Glove induced on just 6 billion of words (glove6B) scores better than sg100B on both

[20] This asymmetry is present despite adjusting weights (to account for the unequal number of datasets) and applying range normalization (to neutralise the conservative bias of harmonic mean).

Table 7 Computational models of relatedness and their ranks with respect to their performance on three specialised datasets: thematic relations (*themrel*), evocation (*evoc*) and Simlex-999 (*simlex*) datasets

Model	Type	*Har-mean*	*Themrel*	*Evoc*	*Simlex*
RWSGwn+glo	a	**1**	**2**	**1**	13
sp-sparse+sg	a	**2**	14	13	4
RWSGwn+sg	a	**3**	11	7	10
RWSGwn+nnse	a	4	13	8	9
RWSGwn	rw	5	15	9	11
sp-sparse+glo	a	6	**3**	**2**	19
ling-svds+sg	a	7	7	10	15
ling-svds+glo	a	8	5	**3**	21
glove840B	d	9	4	4	22
sg100B	d	10	8	11	18
sp-sparse+nnse	a	11	12	19	**1**
ling-sparse+glo	a	12	20	16	**2**
ling-svds+nnse	a	13	9	18	12
glove6B	d	14	6	5	25
glove42B	d	15	**1**	12	24
ddNNSE2500	d	16	10	20	14
sg-window2	d	17	21	14	20
sg-window5	d	18	19	6	27
sp-sparse	l	19	22	22	8
sg-deps	l	20	23	23	17
huang100	d	21	17	15	28
docNNSE300	d	22	16	17	29
ddNNSE300	d	23	25	25	26
CW	d	24	18	21	30
ling-svds	kb	25	28	26	**3**
hso	kb	26	26	28	16
polyen	d	27	24	24	31
turian100	d	28	29	27	32
depNNSE300	l	29	30	29	23
ling-sparse+sg	a	30	31	30	6
ling-sparse	kb	31	32	32	7
ling-sparse+nnse	a	32	33	31	5
hpca100	d	33	27	33	33

The order of models in the table is by weighted harmonic mean of the results on the three datasets. Bold font indicates three best results obtained on each dataset

Type symbols:

kb = knowledge-based models

d = vanilla distributional models

l = linguistically informed distributional vectors (dependency-based and symmetric patterns)

a = concatenated vectors

rw = RWSGwn

Table 8 Spearman correlation between the scores obtained by models at each of the specialized dataset and the sizes of their training corpora, (a) considering only models trained on raw corpora, (b) including linguistically informed distributional models

	Themrel	*Evoc*	*Simlex*
(a) Vanilla distributional models	0.78	0.77	0.73
(b) All distributional models	0.55	0.58	0.70

thematic datasets. As another example, vectors induced by Collobert and Weston (CW) on a small corpus of 0.67 billion tokens outperform another implementation of same network architecture (polyen) trained on 1.8 billion words.

To investigate how the size of the corpus used to train a model impacts its performance on each of the specialised datasets, we defined two groups of models: (a) vanilla distributional vectors (based on simple word-count; see Sect. 4.2) and (b) all co-occurrence based models, i.e. all vanilla distributional models plus dependency-based vectors and symmetric patterns (Sect. 4.3). For each dataset, we calculated Spearman correlation between the list of scores obtained on that dataset by models from group (a), and the list of their training set sizes. The same procedure was performed using models from group (b). The results are presented in Table 8.

The results in Table 8 give rise to several observations. First, within group (a), the impact of training corpus size is higher for the thematic datasets than for the taxonomic dataset (*simlex*). Second, the correlation drops for group (b). This is not surprising, because group (b) includes models that use syntactic information, which have been shown to improve the ability to capture taxonomic relations (Agirre et al. 2009; Levy and Goldberg 2014; Hill et al. 2015; Schwartz et al. 2015). Even when trained on relatively small corpora, these models perform "abnormally" well on *simlex*, diluting the relationship between training set size and the performance. On the other hand, since the specialisation in recognizing taxonomies impairs on the ability to detect thematic relations, these linguistically enriched models perform "abnormally" poorly on the thematic datasets, which also mitigates the impact of training data size.

More intriguingly, however, the drop in correlation observed for the thematic datasets is much more radical than for *simlex*. This might mean that a small improvement on the taxonomic benchmark comes at the price of a huge performance loss on the thematic dimension. Although at this stage this discussion is of a speculative nature, it would be worthwhile to further investigate and verify our results. A better insight into the interaction between the ability to capture taxonomic and thematic relations may facilitate more accurate choices of models to be applied for specific NLP tasks.

Lastly, we notice that the correlations obtained with respect to *themrel* and *evoc*—quite different from correlations calculated for *similex*—are very similar to each other, both within group (a) and (b). We interpret it as an additional premise to support our claim that the thematic datasets consistently capture a meaningful dimension of semantic relatedness.

7.3 Most successful embeddings

Although combined models dominate the top positions in all types of rankings (both by the harmonic mean and by performance on specialized datasets), the winning combinations tend to be different for each type of semantic relatedness. Below we identify specific embeddings that are most successful at (a) balancing the ability to recognize taxonomic and thematic relations, and (b) capturing thematic relations:

(a) Four of the five best balanced models are combinations including RWSGwn or RWSGwn itself. The feature that distinguishes the RWSGwn from other models evaluated in our study is the use of gloss relations. Gloss relations are links between words in synset S and synsets representing words in the gloss of S. A model using gloss relations receives explicit information about thematic relations (for example, *hospital* will be linked to *patients* and *treatment*, because its gloss is: *a health facility where patients receive treatment*). Our results indicate that combining taxonomic information from hierarchical links in WordNet with the direct thematic information from gloss annotations helps to achieve a reasonable trade-off between capturing two incompatible types of relatedness.

(b) The top five performers on both *themrel* and *evoc* are combinations of GloVe vectors or GloVe itself. This consistent advantage of GloVe is only observed on thematic datasets, not on *simlex* or when comparing harmonic mean values. This indicates that the ability of GloVe to capture global corpus statistics may facilitate recognizing thematic aspect of relatedness.

7.4 The harmonic mean and USF Free Association Norms

In calculating the weighted harmonic mean values shown in Tables 6 and 7, we have not included results obtained on USF Free Association Norms because we did not have any pre-existing knowledge about the proportion of taxonomic vs. thematic relations captured in this dataset. Since the purpose of calculating the harmonic mean was to find the best-balanced model, we applied weights so as to find a middle point between scores obtained on the two thematic datasets on one side, and a single taxonomic benchmark on the other. We did not have enough information to determine the weight that should be assigned to USF dataset, because its word pairs were spontaneously generated by participants, and the type of relatedness between a cue and the response was not controlled in any way. Studies in cognitive science (Hutchison 2003; Golonka and Estes 2009) confirm that people use both taxonomic and thematic relations to categorize concepts, but there is no simple answer as to which type is used more often; rather, it is suggested that people's preferences in that regard are influenced by a variety of personal, situational and even cultural factors. Thus, we could expect *usf* to be a mixed dataset but make no assumptions as to whether it is biased toward either type of relatedness.

To shed some light onto the nature of relations gathered in the USF database, we compared ranks by performance on *usf* with ranks by the harmonic mean obtained in Experiment 2. Table 9 shows that the order by scores obtained on *usf* is strikingly similar to the rank by harmonic mean, which suggests that the USF database contains a balanced mixture of taxonomic and non-taxonomic relations. It,

Table 9 Rank by weighted harmonic mean of normalised results (*har-mean*) as compared to the rank by performance on USF Free Association Norms dataset (*usf*)

Model	Model type	Rank by *har-mean*	Rank by *usf*
RWSGwn+glo	a	1	1
sp-sparse+sg	a	2	11
RWSGwn+sg	a	3	2
RWSGwn+nnse	a	4	3
RWSGwn	rw	5	4
sp-sparse+glo	a	6	6
ling-svds+sg	a	7	5
ling-svds+glo	a	8	8
glove840B	d	9	9
sg100B	d	10	7

Type symbols:

kb = knowledge-based models

d = vanilla distributional models

l = linguistically informed distributional vectors (dependency-based and symmetric patterns)

a = concatenated vectors

rw = RWSGwn

therefore, could be a useful resource for evaluating general purpose models; rather than assessing computational measures of relatedness against several specialised datasets and taking a mean result, it may be simpler to use a single gold standard.

8 Conclusions

This paper focuses on differences between two aspects of semantic relations, placing them in the context of evaluating computational models of relatedness. In particular, we highlight the importance of non-taxonomic relations. Traditionally, NLP literature recognizes taxonomic relations as a subset of general semantic relatedness, but relations beyond this subset are poorly defined.

We propose taking advantage of research advances in cognitive psychology and adopting the concept of thematic relatedness, which is well defined, crisply separated from taxonomic relatedness, and proven to play crucial role in human cognition. Improved conceptualization of the non-taxonomic portion of semantic relatedness may facilitate NLP research on this important type of connection between words/concepts.

The duality of semantic relations has been also investigated by linguists and semioticians. Although considered from very different perspectives, the picture emerging from the research across the domains of natural language processing, cognitive psychology and linguistics is surprisingly consistent: both the internal, feature-based taxonomic relations, and the external, contextual thematic relations are equally important for cognitive and linguistic processes, and both are continually used by humans for organizing concepts or constructing and understanding linguistic

messages. One thread that runs through these studies is about the inherent difficulty in defining and interpreting thematic (non-classical, contiguity) relations, which are context-dependent and somewhat subjective. A related common theme is that Western culture and education system emphasize taxonomic relations, focusing on objects and attributes, and discouraging thinking in terms of contexts and complementarity. As a result, thematic relations tend to be neglected or misunderstood. (cf. Jakobson 1956; Morris and Hirst 2004, 2006; Boyd-Graber et al. 2006; McRae and Boisvert 1998; Lin and Murphy 2001; Estes et al. 2011).

Hoping to contribute to a better apprehension of the concept of thematic relations, we demonstrate how this conceptualization can be applied in the evaluation of computational models of semantic relatedness. We identify two thematic datasets that so far have not been used by researchers working with word representations, prepare them for use as evaluation gold standards, and verify their ability to selectively capture non-taxonomic relations. This form of verification has limitations, as non-taxonomic relations are a broader concept than thematic relations; however, the motivation of researchers who designed the two datasets, as well as the procedures taken during collecting human ratings, indicate that the type of semantic relatedness measured by these datasets is specifically thematic relatedness. We recommend them as thematic relatedness benchmarks that complement the well-known taxonomic benchmark, Simlex-999.

Following the assumption that different NLP tasks may target different aspects of semantic relatedness, we use the thematic evaluation datasets and Simlex-999 to analyse the performance of an extensive sample of computational models of semantic relatedness and identify models that are best at each dimension of relatedness. Acknowledging that natural language users manipulate both kinds of relations, and to meet the requirements of applications that need to model that bipolar structure of language, we propose using harmonic mean as a way of assessing the level of balance in capturing the two aspects.

We experiment with concatenated vectors, exploring several novel combinations. Our evaluation provides evidence that combining multiple sources of information brings about a better balance in recognizing taxonomic and thematic relations, as measured in terms of the harmonic mean. Interestingly, concatenated vectors also obtain the highest scores on each of the specialised datasets, albeit different combinations are required for each dataset. We find that combinations including RWSGwn (a model utilizing WordNet with gloss relations) yield the best-balanced systems, while combinations including GloVe (a distributed neural language model that directly encodes global corpus statistics) are most successful at capturing thematic relations.

The last contribution of this paper is verification of USF Free Association Norms dataset, which is not typically employed for evaluation of computational measures of relatedness. USF Free Association Norms dataset is a huge collection (over 60,000 normed cue-response pairs), and has been supplied with rich metadata that comprise additional resources available for researchers interested in specific linguistic tasks. Our results suggest that the USF dataset may be especially useful for evaluating general purpose models, as it seems to cover taxonomic and thematic relations in balanced proportions.

Acknowledgements The work was partly supported by the ADAPT Centre which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

A Supplemental results

See Tables 10, 11 and Fig. 4.

Table 10 Performance scores of vanilla distributional and knowledge-based models evaluated against subsets of datasets (noun-noun and verb-verb pairs only)

	Raw results				Normalized results			
Dataset Words total	Themrel 976	Evoc 6,629	Usf 33,808	Simlex 888	Themrel 976	Evoc 6629	Usf 33,808	Simlex 888
Distributional								
CW	0.15	0.20	0.22	0.27	0.46	0.69	0.45	0.30
turian100	0.06	0.10	0.10	0.21	0.08	0.14	0.00	0.16
polyen	0.12	0.16	0.21	0.25	0.32	0.45	0.39	0.24
hpca100	0.09	0.07	0.14	0.15	0.21	0.00	0.15	0.00
huang100	0.18	0.20	0.24	0.28	0.62	0.70	0.51	0.33
glove6B	0.24	0.24	0.34	0.33	0.88	0.90	0.88	0.45
glove42B	0.27	0.20	0.35	0.34	1.00	0.70	0.91	0.48
glove840B	0.25	0.26	0.38	0.37	0.95	1.00	1.00	0.54
sg100B	0.23	0.23	0.37	0.42	0.83	0.87	0.97	0.66
docNNSE300	0.18	0.16	0.26	0.25	0.60	0.50	0.59	0.25
sg-window5	0.16	0.25	0.33	0.35	0.50	0.95	0.82	0.49
sg-window2	0.14	0.21	0.31	0.39	0.41	0.76	0.76	0.59
Knowledge based								
ling-svd	0.08	0.08	0.23	0.54	0.17	0.02	0.47	0.96
ling-sparse	0.05	0.11	0.24	0.55	0.00	0.19	0.50	0.98
jc	0.08	0.10	0.21	0.56	0.16	0.13	0.42	1.00
lin	0.06	0.09	0.23	0.55	0.04	0.11	0.49	0.96
res	0.05	0.08	0.23	0.47	0.02	0.05	0.48	0.78
lc	0.07	0.09	0.21	0.55	0.10	0.11	0.40	0.97
path	0.07	0.09	0.20	0.52	0.10	0.11	0.37	0.90
wup	0.07	0.09	0.23	0.52	0.10	0.09	0.49	0.90
hso	0.10	0.10	0.25	0.46	0.25	0.15	0.55	0.75

"Words total" is the number of word pairs in a dataset after excluding pairs not supported by WNM

Table 11 Welch's t-test results (unequal variances) of comparing mean scores obtained by a sample of vanilla distributional models (d) and a sample of knowledge-based models (kb) evaluated against pruned versions of four evaluation datasets (noun-noun and verb-verb pairs only)

	Themrel*	Evoc*	Usf**	Simlex*
H_{alt}	$\overline{d} > \overline{kb}$	$\overline{d} > \overline{kb}$	$\overline{d} \neq \overline{kb}$	$\overline{d} < \overline{kb}$
t-statistics	5.191	5.742	1.607	8.407
p-value	0.00009	0.00007	0.136	0.0000001
H_0 rejected?	Yes	Yes	No	Yes

Null hypothesis assumes no difference between sample means

*One-tailed Student's t-test assuming equal variances, unequal sample size

**Two-tailed Welch's t-test assuming unequal variances, unequal sample size

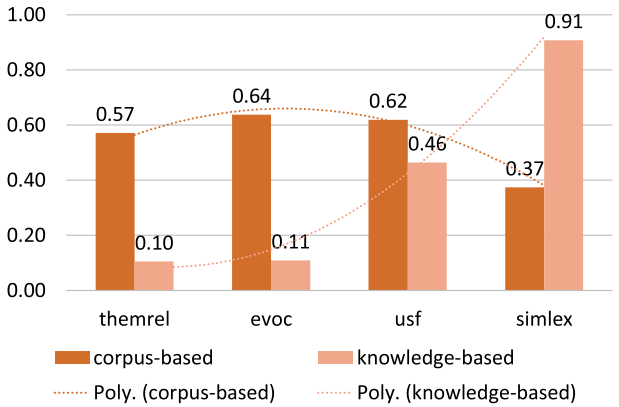

Fig. 4 Average of performance scores obtained by distributional and knowledge-based models (including graph-based WordNet-based measures) on pruned versions of thematic relations norms, evocation, free association norms and Simlex-999. Polynomial trendlines (*Poly.*) are added to accentuate performance patterns

References

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

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the seventeenth conference on computational natural language learning*, Sofia, Bulgaria, pp. 183 – 192.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Boyd-Graber, J., Fellbaum, C., Osherson, D., & Schapire, R. (2006). Adding dense, weighted, connections to WordNet. In *Proceedings of the global WordNet conference*, Jeju, South Korea, pp. 29 – 36.

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. Tech. Rep. CSRG-390, Department of Computer Science, University of Toronto, Toronto, Canada.

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

Collobert R., & Weston J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International conference on machine learning*, ICML.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *CoRR*. arXiv:1103.0398.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. *Psychology of Learning and Motivation: Advances in Research and Theory*, *54*, 249–294.

Faruqui M., Dyer C. (2015). Non-distributional word vector representations. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 2: Short Papers), Association for Computational Linguistics, Beijing, China, pp. 464–469.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*., Language: Speech and communication Cambridge: MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web*, ACM, New York, NY, USA, WWW '01, pp. 406–414.

Finlayson, M. A. (2015). MIT Java Wordnet Interface (JWI) user's guide. *Version*, *2*(4).

Goikoetxea, J., Soroa, A., & Agirre, E. (2015). Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, Association for Computational Linguistics, Denver, Colorado, pp. 1434–1439.

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1454–1464.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2–3), 146–162.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305–332). Cambridge: MIT Press.

Hoyer, P. O. (2002). Nonnegative sparse coding. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing* (pp. 557–565).

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics: Long papers*—Volume 1, Association for Computational Linguistics, pp. 873–882.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813.

Jackson, R. L., Hoffman, P., Pobric, G., & Ralph, M. A. L. (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral Cortex*, *25*(11), 4319–4333.

Jakobson, R. (1956). Two aspects of language and two types of aphasic disturbances. In R. Jakobson, & M. Halle (Eds.), *Fundamentals of language*. The Hague & Paris: Mouton.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of international conference on research in computational linguistics (ROCLING X)*, Taiwan, pp. 19–33.

Jouravlev, O., & McRae, K. (2015). Thematic relatedness production norms for 100 object concepts. *Behavior Research Methods*, *48*(4), 1349–1357.

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge: The MIT Press.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Langone, H., Haskell, B. R., & Miller, G. A. (2004). *Annotating wordnet*. DTIC Document: Technical report.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge: MIT Press.

Lebret, R., & Collobert, R. (2014). Word embeddings through hellinger PCA. In *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics,* Association for Computational Linguistics, Gothenburg, Sweden, pp. 482–490.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics* (Vol. 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, pp. 302–308.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the fifteenth international conference on machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '98, pp. 296–304.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*(1), 3.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 558–572.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, *3*(6), 1417–1429.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*. arXiv:1301.3781.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 1081–1088). Curran Associates, Inc.

Moldovan, D., & Novischi, A. (2004). Word sense disambiguation of WordNet glosses. *Computer Speech & Language*, *18*(3), 301–317.

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

Morris J., & Hirst G. (2006). The subjectivity of lexical cohesion in text. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.) *Computing attitude and affect in text: Theory and applications.* The information retrieval series (Vol. 20, pp. 41–47). Springer, Dordrecht.

Murphy, B., Talukdar, P. P., & Mitchell, T. M. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. In M. Kay, & C. Boitet (Eds.), *COLING* (pp. 1933–1950). Mumbai: Indian Institute of Technology Bombay.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, Association for Computational Linguistics, pp. 38–41.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

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

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 109–117.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on artificial intelligence*—Volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI'95, pp. 448–453.

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

Schwartz, R., Reichart, R., & Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the nineteenth conference on computational natural language learning*, Association for Computational Linguistics, Beijing, China, pp. 258–267.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st national conference on artificial intelligence*—Volume 2, AAAI Press, Boston, Massachusetts, AAAI'06, pp. 1419–1424.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 384–394.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '94, pp. 133–138.

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