COVER: a linguistic resource combining common sense and lexicographic information

Noname manuscript No.
(will be inserted by the editor)

COVER: a Linguistic Resource Combining Common Sense and Lexicographic Information

Enrico Mensa · Daniele P. Radicioni · Antonio Lieto

Received: date / Accepted: date

Abstract Lexical resources are fundamental to tackle many tasks that are central to present and prospective research in Text Mining, Information Retrieval, and connected to Natural Language Processing. In this article we introduce COVER, a novel lexical resource, along with COVERAGE, the algorithm devised to build it. In order to describe concepts, COVER proposes a compact vectorial representation that combines the lexicographic precision characterizing BabelNet and the rich common-sense knowledge featuring ConceptNet. We propose COVER as a reliable and mature resource, that has been employed in as diverse tasks as conceptual categorization, keywords extraction, and conceptual similarity. The experimental assessment is performed on the last task: we report and discuss the obtained results, pointing out future improvements. We conclude that COVER can be directly exploited to build applications, and coupled with existing resources, as well.

Keywords Lexical Resources · Lexical Semantics · Common Sense Knowledge · Vector Representation · Concept Similarity · NLP

Computer Science Department, University of Turin
E-mail: {mensa,radicion,lieto}@di.unito.it

1 Introduction

The growth of the Web and the tremendous spread of social networks [12] exert a strong pressure on computational linguistics to refine methods and approaches to improve applications in areas as diverse as documents categorization [73], conceptual categorization [43], keywords extraction [46], question answering [25], text summarization [28], and many others. The role of linguistic resources —mostly those concerned with lexical semantics— has been herein central: in the last decades, the success in several tasks such as word sense disambiguation has been strongly related to the development of lexical resources [51,53,57]. The same holds for specialized forms of semantic analysis and interpretation, such as sentiment analysis, where systems' efficacy [11] has been accompanied by the release of specialized lexical resources and corpora (e.g., [5,47,18]). Finally, in the last few years the creation of multilingual and parallel resources [21,58] further strengthened the link between lexical resources and successful NLP applications [16,24,56].

In order to provide artificial systems with human-level competence in understanding text documents (which is known to be an AI-complete task [82,32, 41]), one chief component is basically missing from existing resources, with the notable exception of ConceptNet [27]: that is, *common-sense*. Common-sense is assumed to be a widely accessible and elementary form of knowledge [55], whose main traits can be encoded as prototypical knowledge [69]. For example, if we consider the concept *water*, the common-sense knowledge related to this concept is that water, typically, occurs in liquid state and that it is usually a colorless, odorless and tasteless fluid.[1] This is a relevant piece of information, since in many settings artificial agents need to complement more structured information (such as, e.g., about the chemical composition or taxonomic information) with common-sense aspects. However, although ConceptNet is suited to structurally represent common-sense information related to typicality, it cannot be directly connected to further resources due to the fact that it disregards the conceptual anchoring issue (more on this topic later on). Other well known semantic resources, such as DBpedia [2] and the ontological resource Cyc [34], are *de facto* not able to do represent common-sense information. In DBpedia, such information is scattered in textual descriptions (e.g., in the abstracts) rather than being available in a structured, formal and accessible way. For instance, the *fork* entity can be categorized as an object, whilst there is no structured information about its typical usage, places where forks can be found, entities that frequently are found together with forks, *etc.*. As a consequence, DBpedia provides poor results when tested on queries involving common-sense knowledge [37]. Cyc is one of the largest ontologies available, and one of the biggest attempts to build common-sense knowledge bases. De-

[1] "When people communicate with each other, they rely on shared background knowledge to understand each other: knowledge about the way objects relate to each other in the world, people's goals in their daily lives, the emotional content of events or situations. This 'taken for granted' information is what we call common sense – obvious things people normally know and usually leave unstated" [12, p.15].

spite this premise, however, such resource (at least in its publicly available version, OpenCyc) does not represent common-sense information. Similar to DBpedia, in fact, when tested on common-sense queries [36,37], systems built on top of the OpenCyc ontology obtain poor results.[2]

In this work we introduce the lexical resource COVER (so named after 'COmmon-sense VEctorial Representation'), which we propose as a helpful resource to semantically elaborate text documents. COVER is built by merging BabelNet [58], NASARI [9] and ConceptNet [27] with the aim at combining, in a synthetic and cognitively grounded way, lexicographic precision and common-sense aspects. The knowledge representation adopted in COVER allows a uniform access to concepts via BabelNet synset IDs; it consists of a vector-based semantic representation which is also compliant with the Conceptual Spaces, a geometric framework for common-sense knowledge representation and reasoning [23].

Different from most popular vectorial resources that rely on Distributional Semantics, representing hundreds of opaque distributional features (in particular for resources using latent semantic indexing), COVER provides the represented elements with a reduced number of cognitively salient dimensions and, as illustrated in the following, it allows building applications that obtain interesting results in a number of tasks.

2 Related Work

In the last few years different methodologies and systems for the construction of unified lexical and semantic resources have been proposed, as portrayed in Figure 1. In particular, one clear trend has recently emerged: besides resources that have been built either based on manual annotation (such as WordNet [51] and FrameNet [3]) or in automatic fashion (such as BabelNet [59]), many efforts have been spent in building vector representations that are known as distributional semantics models or word embeddings.

2.1 Vector Representations

Let us start from the recent approaches that rely upon vector representations: in this setting, one major assumption is that words that occur in similar contexts tend to purport similar meanings [26]; this principle seems to be compatible with some mechanisms of language acquisition that are based on similarity judgments [83]. Word meanings are herein represented as dense unit vectors of real numbers over a continuous, high-dimensional Euclidean space, where word similarity and relatedness can be interpreted as a metric. Four

[2] The representational limitation of this ontological resource has also led to the development of hybrid knowledge representation systems, such as, e.g., Dual-PECCS [43], that adopts OpenCyc to encode taxonomic information and resorts to different integrated frameworks the task of representing common-sense knowledge.

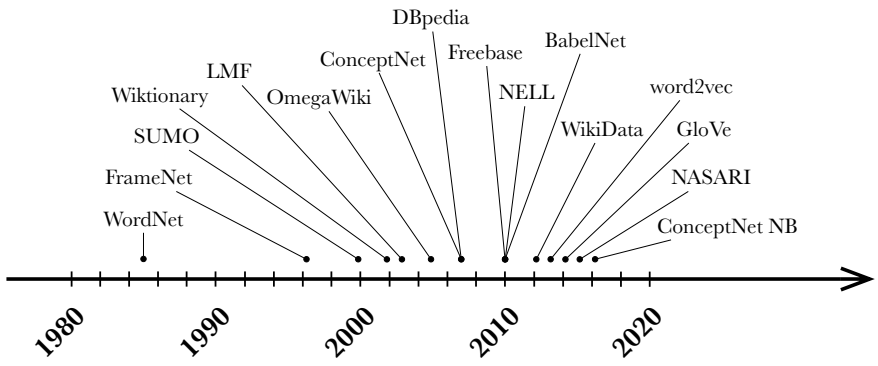

Fig. 1 Mapping on the timeline of some of the most relevant linguistic resources proposed in the last decades.

of the most popular embeddings are word2vec [50], GloVe [64], NASARI [9] and ConceptNet Numberbatch [74]. The word2vec models and the associated off the shelf word embeddings result from a training over 100 billion words from the Google News through continuous skip-grams. The authors of this work exploit simple — compared to either feedforward or recurrent network models — model architectures and illustrate how to train high quality word vectors from huge data sets. While word2vec is commonly acknowledged to be a *predictive* model, GloVe [64] is instead a *count based* model (more on this distinction can be found in [4]). In count based models, model vectors are learned by applying dimensionality reduction techniques to the co-occurrence counts matrix; in particular, GloVe embeddings have been acquired through a training on 840 billion words from the Common Crawl dataset.[3] As regards as the more recent ConceptNet Numberbatch [74, 75], it has been built through an ensemble method combining the embeddings produced by GloVe and word2vec with the structured knowledge from the semantic networks ConceptNet [76] and PPDB [22]. The authors employ here locally-linear interpolation between GloVe and word2vec, and also propose adopting ConceptNet as knowledge source for retrofitting distributional semantics with structured knowledge [19].

Some other related works are concerned with the extraction of Conceptual Spaces representations. Conceptual Spaces are a cognitively-inspired representational framework assuming that conceptual knowledge in human and artificial systems, is ultimately represented and used for intelligent tasks in small-scale geometric spaces (i.e., in a specific characterization of vector-based representations). In such framework, knowledge is represented as compact set of quality dimensions and a geometric or topological interpretation is associated to each quality dimension (we refer to [23] for the details on the framework). Existing approaches, for example, try to induce Conceptual Spaces based on distributional semantics by directly accessing huge amounts of tex-

[3] `http://commoncrawl.org`.

tual documents to extract the multidimensional feature vectors that describe the Conceptual Spaces. In particular, the work by [17] tries to learn a different vector space representation for each semantic type (e.g., movies), given a textual description of the entities in that domain (e.g., movie reviews). Specifically, in the mentioned work, the authors use multi-dimensional scaling (MDS) to construct the space and identify directions corresponding to salient properties of the considered domain in a *post-hoc* analysis. A similar (though more limited) approach has been recently undertaken in [38], consisting of automatically extracting some basic and perceptually prominent feature values, such as for the dimensions Shape, Size, Location, etc..

Since term meanings are represented as points, vectors and regions in a Euclidean space, CSs and word embeddings can be considered to some extent as cognate representations. However, word embeddings also differ in at least two crucial ways that limit their usefulness for applications in knowledge representation, e.g., in automatically dealing with inconsistencies. First, word embedding models are mainly aimed at modelling *similarity* (and notions such as *analogy*, like in the Latent Relational Analysis approach by [78]), and are not designed to provide a geometric representation of conceptual information (e.g., by representing concepts as convex regions where prototypical effects are naturally modelled). Moreover, the dimensions of a word embedding space are not directly interpretable in that the meaning of the features is not directly accessible, while quality dimensions in Conceptual Spaces directly reflect salient cognitive properties of the underlying domain. This fact has direct impact on the *explanatory* capacity of word embeddings: the similarity between two entities is assessed based on the closeness of their vector representations in a multidimensional space according to some given metrics. *Retrofitting* techniques have been proposed to refine vector space representations by borrowing information from semantic *lexica* [19]. However, these can be used rather to smartly find out terms with closer vector representation, rather than to introduce information on features, functions and roles, which would explain why and in how far two entities are similar or related.

The vector representations conveyed by word embeddings have been adopted in systems that exhibit good (impressive, in some cases [75, 7]) agreement with human judgment and they can be applied in some specific tasks such as analogical reasoning; however, no justification based on properties/relations is allowed in this setting. Conversely, no wide coverage lexical resource has been so far carried out that is fully compliant to Conceptual Spaces, also due to the fact that Conceptual Spaces have been designed to grasp mainly perceptual qualities, and they can be hardly generalized to any arbitrary domain.

2.2 Annotation Based Representations

Another broad class of lexical resources includes a heterogeneous set of works that can be arranged into hand-crafted resources — created either by expert annotators, such as WordNet [51], FrameNet [3] and VerbNet [35], or through

collaborative initiatives, such as ConceptNet [27] —; and resources that have been built by automatically combining the above ones, like in the case of BabelNet [59].

WordNet (WN) is a lexical database for the English language. It has been the first and the most influential resource in the field of lexical semantics; its hierarchies are to date at the base of other resources, and it is has been used in various and diverse sorts of applications, such as, e.g., to compute supersense tagging [13] and several tree-based similarity metrics [63]. Different from traditional dictionaries — organizing terms alphabetically, thus possibly scattering senses — WN relies on the idea of grouping terms into synonyms sets (called *synsets*), that are equipped with short definitions and usage examples. Such sets are represented as nodes of a large semantic network, whose edges express semantic relations among synset elements (such as hyponymy, hypernymy, antonymy, meronymy, holonymy). BabelNet is a wide-coverage multilingual semantic network resulting from the integration of lexicographic and encyclopedic knowledge from WordNet and Wikipedia, respectively; it extends the constructive rationale of WN — and as such it is also based on sets of synonyms, the Babel synsets — through the structure of Wikipedia composed of redirect pages, disambiguation pages, internal links, inter-language links, and categorical information. More on the algorithm used to build BabelNet can be found in [59].

None of the mentioned proposals addresses the issue of integrating resources and extracting information to the ends of providing common-sense conceptual representations, also provided with a thorough conceptual anchoring. The rationale underlying COVER is to extract the conceptual information hosted in BabelNet (and its vectorial counterpart, NASARI [9]) and to exploit the relations in ConceptNet so to rearrange BabelNet concepts into a semantic network enriched with ConceptNet relations. Differently from the surveyed works, however, this is done by leveraging the lexical-semantic interface provided by such resources. In the next Section we illustrate our strategy in building our resource.

3 The COVERAGE Algorithm and the COVER Lexical Resource

Before introducing COVER, we illustrate COVERAGE (that stands for COVER Automatic GEnerator), the algorithm designed to build COVER. The goal of the COVERAGE algorithm is to create a collection of semantic vectors, one for each concept c provided as input. Each obtained vector $\vec{c}$ contains common-sense information about the input concept, and it is encoded as a set of semantic dimensions D. More precisely, each dimension (e.g., HasPart or UsedFor) contains a set of concepts that constitute the values filling that dimension for the concept c. The adopted algorithm relies upon two well-known semantic resources, that are NASARI [9] and ConceptNet [76].

```
NASARI unified vector:
bn:00000001n    -NA-  bn:00021248n_73.1 bn:00005513n_14.48 ...

NASARI embedded vector:
bn:00000001n    -NA-    0.02738748  0.00093856  0.0698559 ...
```

Fig. 2 The NASARI and NASARIE vectors for the `bn:00000001` concept. The first element of the vector is the BSI, that identifies the concept associated with the vector; the second one is the Wikititle (an unnamed concept is illustrated in this case, `-NA-`); the remaining elements are either BSIs enriched with their weight in the NASARI unfied vector, or float numbers in the NASARIE vector.

3.1 Employed Resources

NASARI. NASARI is a set of distributional semantic vectors, each one providing distributional information regarding a concept, identified through a BabelNet synset ID (hereafter also BSI). We employ two out of the three available NASARI versions:

- NASARI unified: each vector contains a weighted list of other concepts (also identified by BSIs) semantically close to the concept being represented by the current vector;
- NASARI embedded (referred to as NASARIE from now on): each vector defines a dense vector in a 300-dimensions space. All the NASARIE vectors share the same semantic space, so that these representations can be used to compute semantic distances between any two such vectors.

The two different representations (NASARI and NASARIE vectors) for the same concept are illustrated in Figure 2. The NASARI vectors are used as sense inventory and provide a connection between the *term* and the *sense* level. Because we rely on BSIs in order to identify the different senses, and because BabelNet is a multi-language resource, it follows that also COVER is a multilingual resource.

ConceptNet. ConceptNet is a semantic network, where nodes represent words and phrases connected through a large set of relationships. We chose to extract the information from ConceptNet because it is mainly constituted by commonsense knowledge, as illustrated by the dump provided in Figure 3. However, since this resource does not provide a clear semantic grounding, nodes herein conflate all possible senses. Let us briefly elaborate on the main differences between ConceptNet and NASARI, by comparing their limitations and merits in order to introduce the main axes that drove the design of COVER.

Motivation for Merging NASARI and ConceptNet

As it emerges from the above discussion, NASARI contains a set of conceptually grounded vectors. Each such vector is constituted by concepts that are

Fig. 3 Representation of the node *table* in ConceptNet.

semantically proximal, leaving unspecified the nature of their semantic connection. For instance, the vector describing *table* ("A piece of furniture having a smooth flat top that is usually supported by one or more vertical legs", identified as BN:00075813N) may be related to (the BSIs corresponding to) *furniture*, *leg*, *kitchen* and so forth, but it provides no further information on *why* and *how* each of these entities is related to *table*. On the other side, ConceptNet is built upon relationships, but it doesn't provide any conceptual grounding to the involved nodes. Specifically, ConceptNet nodes are not concepts but lexical entities (and possibly *compound words*, such as "Something you find inside"). In this sense, ConceptNet offers a much richer and descriptive vocabulary, but at the expense of a reduced 'ontological' and taxonomic precision (no concept identifier is used at all). For example, we have that *table* IsA *forniture*, HAS *legs*, and can be found AtLocation *kitchen*. However, given the absence of a conceptual grounding, the same *table* node will also provide relationships such as *table* IsA *contents*, IsA *counter*, IsA *calendar*, thus resulting in a mixture of relationships regarding all possible senses underlying the given term *table* (please refer to Figure 3).

COVER representation benefits from the rich set of relations from ConceptNet, and from the lexicographic precision proper to (BabelNet and) NASARI. Two main design principles lie at the base of COVER: *i)* the need to make explicit the relationships intervening between a given concept and

```
Exemplar bn:00008010n (bakery, bakeshop)

bn:00008010n_RELATEDTO = [dough, cake, wheatberry, baked goods,
crusty bread, bakehouse, chocolate brownie, breadmaker, buns, produce, shop]

bn:00008010n_ISA = [workplace, sales, outlet, shop]
```

Fig. 4 The vector for the *bakery* concept. The values filling the dimensions RelatedTo and IsA are concepts identifiers (BabelNet synset IDs); for the sake of the readability they have been replaced with their corresponding lexicalization.

those describing it; and *ii)* the need for filling such relations with fully fledged concepts rather than terms/compound words.[4]

3.2 Representation of Lexical Concepts in COVER

The vectors in COVER are defined on a set D of 44 dimensions[5] corresponding to the most salient relationships available in ConceptNet. Each dimension contains a set of values that are concepts themselves, identified through their own BSIs. So a concept c_i has a vector representation $\vec{c_i}$ that is formally defined as

$$\vec{c_i} = [s_1^i, .., s_N^i], \tag{1}$$

where each s_h^i is the set of concepts filling the dimension $d_h \in D$. Each s can contain an arbitrary number of values, or be empty. For instance, the vector bn:00008010n that represents *bakery*, has two dimensions filled (RelatedTo and IsA), and therefore it has two non-empty sets of values (Figure 4).

3.3 Selecting the Sense Inventory: the CloseSt Algorithm

The COVERAGE algorithm takes in input a concept (represented as a BSI) and produces an associated common-sense vector representation. In order to obtain the concepts that are actually fed to the system we start from a set of English terms, in particular, all of the English *nouns* have been retrieved from

[4] Of course, not all information available in ConceptNet can be directly mapped onto BSIs (e.g., the compound word "Something you find inside" has no counterpart in BabelNet/NASARI).

[5] InstanceOf, RelatedTo, IsA, AtLocation, dbpedia/genre, Synonym, DerivedFrom, Causes, UsedFor, MotivatedByGoal, HasSubevent, Antonym, CapableOf, Desires, CausesDesire, PartOf, HasProperty, HasPrerequisite, MadeOf, CompoundDerivedFrom, HasFirstSubevent, dbpedia/field, dbpedia/knownFor, dbpedia/influencedBy, dbpedia/influenced, DefinedAs, HasA, MemberOf, ReceivesAction, SimilarTo, dbpedia/influenced, SymbolOf, HasContext, NotDesires, ObstructedBy, HasLastSubevent, NotUsedFor, NotCapableOf, DesireOf, NotHasProperty, CreatedBy, Attribute, Entails, LocationOfAction, LocatedNear.

the Corpus of Contemporary American English (COCA), which is a corpus covering different genres, such as spoken, fiction, magazines, newspaper, academic.[6] The subsequent step consists of providing each term with the most relevant associated sense(s); this processing is performed by a module implementing the CLOSEST algorithm. It is acknowledged that too fine-grained semantic distinctions may be unnecessary and even detrimental in many tasks [61]: the CLOSEST algorithm accesses BabelNet and produces more coarse-grained (with respect to BabelNet) sense inventories, based on a simple heuristics building on the notions of *availability* and *salience* of words and phrases [80]. Specifically, more central senses are hypothesized —in accordance with their use in spoken and written language— to be more richly represented in encyclopedic resources, to be typically featured by richer and less specific information, and by richer semantic connections with other concepts.

Given an input term t, the algorithm first retrieves the set of senses $S = \{s_1, s_2, \ldots, s_n\}$ that are possibly associated to t: such set is obtained by directly querying NASARI. The output of the algorithm is a result set that is obtained through a process of incremental filtering of S, arranged into two main phases:

1. *LS-Pruning*. Pruning of less salient senses: senses with associated poor information are eliminated. The salience of a given sense is determined by inspecting its NASARI vector;
2. *OL-Pruning*. Pruning of overlapping senses: for each two senses with significant overlap (a function of the number of features shared in the corresponding NASARI vectors), the less salient sense is pruned.

Further details on the CLOSEST algorithm can be found in [39].

Once the sense inventory for each term has been filtered, and a more coarse one has been obtained, the COVERAGE algorithm comes to play.

3.4 The COVERAGE algorithm

The algorithm implemented by COVERAGE can be broken down into two main steps. Given in input a concept c represented by its BabelNet synset ID, the system performs the following operations:

1. **Semantic Extraction**:
   - *Extraction*: all nodes possibly representing c in ConceptNet are retrieved, and all the relevant terms connected to such nodes are triggered and placed in the set of extracted relevant terms T (more about relevance criteria later on).
   - *Concept Identification*: all terms $t \in T$ are disambiguated by equipping each one with a BSI; this step amounts to translating T into the set of relevant extracted concepts C.
2. **Vector Injection**: each concept $c_i \in C$ is injected into its vector representation $\vec{c}$ by exploiting the relationship formerly connecting c_i to c in ConceptNet.

[6] http://corpus.byu.edu/full-text/.

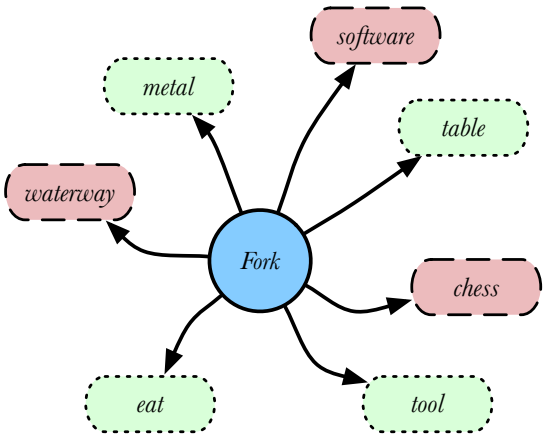

Fig. 5 Each term connected to the ConceptNet node *Fork* is inspected to determine whether it is relevant (dotted contour) or not (dashed contour) for the sense conveyed by the input concept c. While the dotted nodes are relevant because they are referring to *Fork* as the "kitchen utensil" —that is, the sense of c—, the dashed ones refer to *Fork* as the system call for creating processes (*software* node), as the chess move (*chess* node), or as the bifurcation of a watercourse (*waterway* node).

In the next sections we will illustrate the algorithm in detail by following the execution upon the concept $c = \text{BN:00035902N}$, that is *Fork* intended as "the utensil used for eating or serving food".

3.4.1 Semantic Extraction

The Semantic Extraction phase has been designed to build the set C, containing the relevant concepts that will provide the common-sense information for the output vector $\vec{c}$. The first step is the retrieval of the NASARI (unified) vector of c: such task can be performed straightforwardly, thanks to the fact that NASARI is indexed and accessed through BSIs.

The Extraction starts by retrieving all of the ConceptNet nodes that are possibly relevant for c. Because ConceptNet nodes are compound concepts [45] possibly expressed by multi words phrases, we search for all the nodes in ConceptNet that correspond to any term included in either the BabelNet synset or the WordNet synset of c. For example, in the *Fork* case, we look for the nodes *Fork*, *King of utensils*, *Pickle fork*, *Fish fork*, *Dinner fork*, *Chip fork* and *Beef fork* in ConceptNet. All the associations starting from these nodes are collected, and considered as information potentially pertinent to c. However, since we are interested in working at the semantic level, we need to inspect each one of the retrieved associations in order to determine if they are *relevant* to the sense conveyed by c. Figure 5 illustrates the *Fork* node in ConceptNet and its relevant/non relevant connected nodes.

The relevance is assessed by applying two criteria, that are defined as follows.

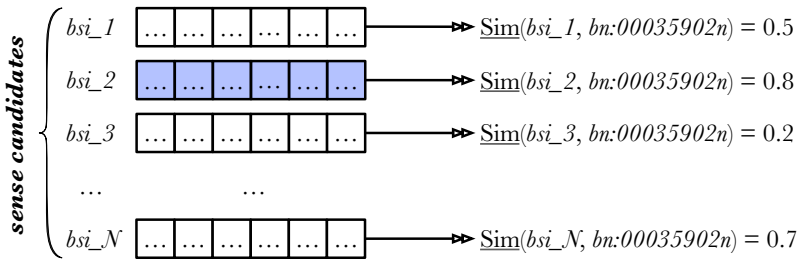

Fig. 6 The similarity between NASARIE candidate vectors and the vector of *Fork* (BN:00035902N) is computed. The highlighted vector is selected, because its similarity with the *Fork* vector obtained the highest score.

Definition 1 (Relevance Criteria) An extracted term t is considered relevant for the concept c if either: *i)* t is included in at least one of the synsets listed in the NASARI vector representation for c; or *ii)* at least β nodes directly connected to t in ConceptNet can be found in the synsets that are part of the NASARI vector representation for c.

The rationale underlying these criteria is explained by the fact that since the NASARI unified vector of c contains concepts (along with their lexicalizations) semantically close to c, the presence of t (first condition) or β terms from its ConceptNet neighborhood (second condition)[7] in such vector guarantees that t is somehow related to c, and it can be thus considered as relevant.

Once the relevance detection is performed, all the relevant terms extracted from all the ConceptNet nodes that we previously collected are put together in the set T. In the *Fork* example, the resulting set would be:

$$T = \{plate,\ tool,\ food,\ utensil,\ silverware,\ table,\ metal\ knife,\ spoon,\ eat\} \quad (2)$$

After having obtained T, that is a set of terms that are guaranteed to be relevant for the sense conveyed by c, the process goes through the *Concept Identification* step. In fact, T still contains lexical elements and not BSIs. A step of word sense disambiguation is thus required in order to convert T into C, by assigning a BSI to each of the terms in T.

The Concept Identification is performed in two different ways, depending on how the term $t_i \in T$ that we are trying to disambiguate was evaluated as relevant during the relevance detection phase. More precisely, if t_i was evaluated as relevant via the first condition (t_i was part of the NASARI vector of c), we automatically have its BSI, thanks to the inner structure of the NASARI vectors (Section 3.1). If, on the other side, t_i was found relevant in virtue of the second condition (Definition 1), we cannot directly retrieve its BSI. In this case, the Concept Identification starts by detecting all the possible meanings of the term. This operation is straightforward: since each BabelNet synset contains all the lexicalizations corresponding to the concept it represents, we retrieve the list of candidate BSIs by selecting those BabelNet synsets that contain

[7] The parameter β has been set to 2 to build the released resource.

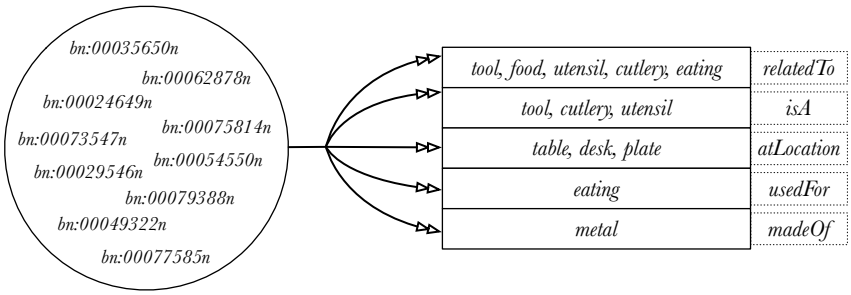

Fig. 7 All the concepts in C are injected into the vector for *Fork*. The concepts identifiers in the vector have been replaced with their lexicalization in order to make the image human readable.

t_i among their lexicalized elements. Subsequently, we retrieve the NASARIE vector associated to each candidate, thus obtaining a set of candidate vectors, one for each BSI possibly appropriate as the meaning of t_i. In either case the selection of the best candidate is performed by exploiting such NASARIE vectors. We first compute the cosine similarity between each candidate vector and the NASARIE vector of c. If the similarity of the most similar vector is greater than a fixed threshold,[8] then the BSI of that vector becomes the meaning of t_i. Figure 6 illustrates this process for the *Fork* example.

Once the Concept Identification is completed, the term t_i is enriched with its BSI and included in the set of the relevant extracted concepts C.

3.4.2 Vector Injection

The second phase of the COVERAGE system consists in the injection of the concepts in C into the vector representation for the input concept c. Each $c_i \in C$ has been retrieved from some node in ConceptNet that was a lexicalization of c, and therefore we have a ConceptNet relationship that connects each c_i to c. Because the COVER vectors have a set of ConceptNet relationships as dimensions (Section 3.2), we just have to properly place each c_i into the dimension corresponding to the relationship that was linking it to c in ConceptNet. Figure 6 illustrates the Vector Injection for the *Fork* example.

The Vector Injection concludes the execution of the COVERAGE system: in the next Section we present some details about the data that has been fed in input to the COVERAGE system and the returned output.

3.5 Building COVER

We now present some features and statistics regarding the computation of COVERAGE, including the size of the lexical base taken as input, some figures on retrieved (and discarded) concepts, and a final quantitative description of the amount of information finally encoded in COVER. In order to

[8] Presently set to 0.6.

Resource	Size
NASARI/NASARIE *vectors*	2,868,176
ConceptNet *assertions*	4,227,874
ConceptNet *nodes*	859,932

Table 1 Information contained in NASARI and ConceptNet, and used as the starting point to build COVER.

obtain the concepts that are actually fed to the COVERAGE algorithm, we start from terms in the Corpus of Contemporary American English and we exploit the CLOSEST algorithm. The CLOSEST system took 27,006 terms in input, and returned 40,816 concepts in output, which were then fed to the COVERAGE system; i.e., some of the terms have been mapped on multiple concepts.

After such preprocessing step, the building of the COVER resource took place. Before the execution of the algorithm, the dataset in input was pruned: 8,979 concepts were dropped as either duplicates (8,867) or because no associated NASARI vector was found (112). Thus 31,837 concepts were fed to the system. The size of the resources employed all throughout this process is reported in Table 1.

As regards as the *Semantic Extraction* phase, overall 4,324,971 terms were extracted from ConceptNet (on average, 135.85 per input concept), but only 42.9% of them (overall 1,856,888) were found relevant. Therefore, the average cardinality of T for each input was 58.32. The concept identification was successful for the 32.61% of such relevant terms, thereby resulting in a total of 605,450 extracted relevant concepts (the average cardinality of the bag of concepts C was then 19.02). We note that roughly two thirds of the concept identification failures were due to the violation of the concept similarity threshold. This threshold is indeed a very sensitive parameter that allows tuning the amount of noise (vs. completeness) featuring the resource: e.g., by setting the similarity threshold to 0.5 instead of 0.6, the average cardinality of C raises to 25.86 (which directly compares with the actual value, 19.02).

As regards as the *Vector Injection* phase, since COVERAGE only loads the ConceptNet relationships that are included into our vectors schema, all the concepts in C were injected into the output vectors. Therefore, the average filling factor (that is, the number of values per concept) corresponds to the average cardinality of C (19.02). This figure was then increased by adding the first 5 elements contained in the NASARI vector for the input concept in its RELATEDTO dimension, bringing the average population of the vectors to 23.97. More precisely, half vectors contain 5 to 20 values, while only 0.5% vectors are filled by less than five values. The most populated dimensions are RELATEDTO, SYNONYM, ISA, HASCONTEXT, ANTONYM, FORMOF and DERIVEDFROM: this distribution closely approaches the distribution of information contained in ConceptNet (Table 2).

The COVERAGE system obtained an empty set C for 4,786 concepts out of the 31,837 provided as input. In such cases, the resulting vectors for

Relationship	Number of associations	% of associations
RelatedTo	1,449,431	51.25%
FormOf	273,560	09.67%
IsA	247,387	08.75%
Synonym	237,772	08.41%
HasContext	177,677	06.28%
DerivedFrom	116,243	04.11%
UsedFor	42,443	01.50%
SimilarTo	29,480	01.04%
AtLocation	28,960	01.02%
CapableOf	26,354	00.93%
HasSubevent	25,896	00.92%
HasPrerequisite	23,493	00.83%
EtymologicallyRelatedTo	20,723	00.73%
Antonym	19,967	00.71%
Causes	17,088	00.60%
HasProperty	13,553	00.48%
PartOf	12,795	00.45%
MotivatedByGoal	9,807	00.35%
ReceivesAction	8,383	00.30%
HasA	7,735	00.27%

Table 2 Distribution of values inside ConceptNet 5.5.0 (only the 20 most populated associations are shown).

such concepts contain exclusively values that were automatically taken from NASARI and injected into the RelatedTo dimension. More in detail, in most failure cases (namely, $4,570$) the system either could not detect any extracted relevant term, or it could not disambiguate any one of the extracted terms. For instance, the input *recantation* produced only *recall* as extracted term. However, the similarity between these two concepts was under the threshold β, therefore, *recall* couldn't be accepted and the C set for *recantation* resulted empty. In the remaining 216 cases, it was not possible to find a ConceptNet node for the input concept. We observed that the vast majority of this concepts contained a dash (e.g., *tete-a-tete*, *god-man*, *choo-choo*). A further improvement would consist in the removal of such dashes in order to detect a suitable ConceptNet node for this kind of inputs.

The COVER resource can be downloaded at the URL `http://ls.di.unito.it/resources/cover/`.

3.6 Applications

The COVER resource has been successfully applied in different tasks, such as the conceptual categorization task, keywords extraction, and for the computation of semantic similarity, both at the word and sense level.

Conceptual categorization. COVER has been used as a knowledge base employed by a system designed to solve the task of conceptual categorization [40, 42, 43, 44]. The task is defined as follows: given a a simple common-sense linguistic description, the corresponding target concept has to be identified. In this setting, a hybrid reasoning system (named DUAL-PECCS, after 'Prototypes and Exemplars-based Conceptual Categorization System') has been devised combining both vector representations and formal ontologies. In particular, DUAL-PECCS is equipped with a hybrid knowledge base composed of heterogeneous representations of the same conceptual entities: that is, the hybrid knowledge base includes prototypes, exemplars and classical representations for the same concept. As regards as the former component of the KB, it is composed by a linguistic resource similar in essence to COVER, although with limited coverage. The whole categorization pipeline implemented by DUAL-PECCS works as follows. The input to the system is a simple linguistic description, like 'The animal that eats bananas', and the expected output is a given category evoked by the description (e.g., the category *monkey*). An algorithm to compute vector distances is executed, that returns an ordering of the concepts that best fit to those in the COVER resource. Then, these results are checked for consistency against the deliberative sub-system, employing standard ontological inference. Interestingly enough, we showed that common-sense descriptions such as that in the example cannot be easily dealt with with ontological inference alone, nor through other standard approaches.

Keywords Extraction. COVER has been used in the keywords extraction task [14]. We investigated a novel approach to keywords extraction that relies on the following assumption: instead of using graph-based models underpinning on terminological information, our system individuates the concepts featuring document content. Their *relevance* as keywords is estimated through their conceptual centrality w.r.t. the concepts in the title. We compared several metrics to compute such relevance: the metrics at stake were based on NASARI (both unified and embedded) vector representations [8], on the COVER representation, and on two further metrics originally conceived to evaluate the coherence in latent topic models [54, 60]. Our experimentation showed that the results obtained through the COVER metrics achieve the highest precision, and competitive accuracy with state-of-art systems [29] on a benchmark dataset [46].

Concept Similarity with Explanation. Additionally, the COVER resource has been used to compute conceptual similarity. One main assumption underlying our approach is that two concepts are similar insofar as they share some values on the same dimension, such as when they are both used for the same ends, they share components, they can be found in the same place(s), etc.. Consequently, our metrics to compute conceptual similarity does not employ WordNet taxonomy and distances between node pairs, such as in [81, 33, 71], nor it depends on information content accounts either, such as in [67, 30], nor it relies on distances between vectors like in embedded representations [10, 74]. Although the system devised does not

```
The similarity between lizard and crocodile is 1.99 because
- they ARE reptile;
- they are RELATEDTO reptile, Caiman, fauna, diapsid.

The similarity between Harry Potter and wizard is 2.50 because
- they are RELATEDTO spell, magic, magician, wand.

The similarity between beach and coast is 2.79 because
- they ARE shore;
- they are semantically SIMILARTO shore, formation;
- they are RELATEDTO shore, coast, weather, seaboard, island, shell,
    wave.

The similarity between sodium chloride and salt is 3.56 because
- they are MADEOF sodium_chloride, ion, crystal;
- they can be found ATLOCATION Shaker_(laboratory), seawater, water,
    nutrient, mine, salt_mine;
- they ARE binary_compound, taste, chemical_compound, Ionic_compound,
    spice, crystal, sodium_chloride, seasoning, inorganic_compound;
- they are USEDFOR seasoning, nutrient;
- they share the same CONTEXT chemistry, inorganic_compound;
- they are SEMANTICALLYOPPOSITE of carbohydrate,
    Swedish_ethyl_acetate_method, vinegar;
- they are PARTOF seawater, sea;
- they are SIMILARTO Sharp_(flour);
- they are SEMANTICALLYSIMILARTO saltiness, sodium_chloride, salinity,
     salt;
- they are RELATEDTO magnesium_lactate, Mevalonic_acid, cholic_acid,
    sulfate, halobacterium, benzoate, sulfonate, monosodium_glutamate,
     Glutaric_acid;
- they are DERIVEDFROM salinity, sodium, chloride, sodium_carbonate.
```

Fig. 8 Some examples of the explanations that can be generated based on the COVER resource; the terms at stake are marked with italic and bold font, while the dimensions are marked with italic font. The similarity values are on a scale from 0.00 to 4.00.

yet achieve state-of-the-art scores (as reported in the next Section), the COVER resource allows to naturally build explanations for the computed similarity by simply enumerating the concepts shared along the dimensions of the vector representation [15], as illustrated in Figure 8. The ability to provide explanations justifying the obtained results is a feature shared by all mentioned applications built on top of COVER; at the best of our knowledge, none of the existing approaches allows to compute such explanation. Further investigations are in progress in order to obtain proper benchmarks on the generated explanations.

We report our evaluation of the COVER resource on the Conceptual Similarity task, that can be thought of as an enabling technology to cope with all the aforementioned applications.

4 Evaluation

The intrinsic evaluation of the completeness and correctness of a lexical resource can be challenging. As testbed to assess COVER, we then opted for an extrinsic evaluation, and considered the conceptual similarity task, which is a long-standing tasks in the lexical semantics field [52,68,81,66]. To these ends, we designed the MeRaLi system, that computes semantic similarity at both sense and word level by specifically relying on COVER. MeRaLi was originally presented in the frame of the Sem-Eval 2017 campaign on Multilingual and Cross-lingual Semantic Word Similarity [48]; we now present a novel experimentation, where the system employs an updated version of the COVER resource.

In this Section we first illustrate the tasks and the similarity metrics implemented by the MeRaLi system; we then introduce the data sets used for testing, and provide the results along with their discussion.

4.1 The Concept Similarity Task

The concept similarity task consists in the estimation of a number that represents the similarity between two proposed concepts.

In our setting, the concept similarity task is actually cast to a vector-comparison problem. In fact, since concepts in COVER are represented as vectors, each one containing other concepts (as depicted in Equation 1), the basic underlying rationale is that two vectors are similar if they share a good amount of information. This criterion to compute conceptual similarity is underpinned by the assumption that two concepts are similar insofar as they share some values on the same dimension, such as when they both share components or properties, inherit from the same superclass, when both entities are capable of performing the same actions, *etc.*. Consequently, our similarity metrics does not employ the WordNet taxonomy and the distances between pairs of nodes, such as in [81,33,71], nor it depends on information content accounts either, such as in [30,67].

Given two input concepts c_i and c_j, after the retrieval of the corresponding COVER vectors $\vec{c}_i$ and $\vec{c}_j$, we compute the similarity by counting, dimension by dimension, the set of values that $\vec{c}_i$ and $\vec{c}_j$ share. Then, the similarity score obtained over each dimension is combined by obtaining an overall similarity score, that is our final output. So, given N dimensions in each vector, the similarity value, $\text{sim}(\vec{c}_i, \vec{c}_j)$, should be ideally computed as:

$$\text{sim}(\vec{c}_i, \vec{c}_j) = \frac{1}{N} \sum_{k=1}^{N} |s_k^i \cap s_k^j|. \tag{3}$$

However, this formulation resulted to be too naïve. In fact, the information available in COVER is not evenly distributed, that is, it may happen that a given dimension is filled with many values (concepts) in the description of a given concept, but the same dimension may be empty in the description of another one. It was hence necessary to refine the above formula to tune the balance between the amount of information available for the concepts at stake: *i)* at the individual dimension level, to balance the number of concepts that characterize the different dimensions; and *ii)* across dimensions, to prevent the computation from being biased by more richly defined concepts (i.e., those with more dimensions filled). Both *desiderata* are satisfied by the Symmetrical Tversky's Ratio Model [31] (which is a symmetrical reformulation for the Tversky's ratio model [79]),

$$\mathrm{sim}(\vec{c}_i, \vec{c}_j) = \frac{1}{N^*} \cdot \sum_{k=1}^{N^*} \frac{|s_k^i \cap s_k^j|}{\beta\,(\alpha a + (1-\alpha)\ b) + |s_k^i \cap s_k^j|} \tag{4}$$

where $|s_k^i \cap s_k^j|$ counts the number of shared concepts that are used as fillers for the dimension d_k in the concept $\vec{c}_i$ and $\vec{c}_j$, respectively; a and b are defined as $a = \min(|s_k^i - s_k^j|, |s_k^j - s_k^i|)$, $b = \max(|s_k^i - s_k^j|, |s_k^j - s_k^i|)$; finally N^* counts the dimensions actually filled with at least two concepts in both vectors. This formula allows tuning the balance between cardinality differences (through the parameter α), and between $|s_k^i \cap s_k^j|$ and $|s_k^i - s_k^j|, |s_k^j - s_k^i|$ (through the parameter β).[9]

4.1.1 Word Similarity

Since some of the data we adopted in the experimentation is actually composed by simple terms (rather than senses), this distinction deserves a brief clarification.

As regards as the computation of the similarity at the words-level, we compute it as the similarity of the closest senses of the words pair; the underlying rationale is that each term works as the context for the other one (e.g., in the pairs ⟨'fork','system call'⟩, and ⟨'fork','river'⟩). In particular, to compute the semantic similarity between a term pair, we adopt a variant of a general disambiguation approach formerly proposed in [62], formulated as follows.

Given: a pair $\langle w_t, C\rangle$, where w_t is the term being disambiguated, and C is the context where w_t occurs, $C = \{w_1, w_2, \ldots, w_n\}$, with $1 \leq t \leq n$; also, each term w_i has m_i possible senses, $s_1^i, s_2^i, \ldots, s_{m_i}^i$.
Find: one of the senses from the set $\{s_1^t, s_2^t, \ldots, s_{m_t}^t\}$ as the most appropriate sense for the target word w_t.

The basic idea is to compute the semantic similarity as a function maximizing the similarity between each two senses (corresponding to the target term and

[9] The parameters α and β were set to .8 and .2 for the experimentation.

to all terms in the context C) by finding the best sense s_h^t disambiguating w_t where h is computed as:

$$h = \underset{m_i=1}{\overset{m_t}{\operatorname{argmax}}} \left[\sum_{w_j \in C, j \neq t} \max_{k=1}^{m_j} \otimes \left(s_i^t, s_k^j \right) \right] \tag{5}$$

where $\otimes$ is implemented by the similarity metrics illustrated in Formula 4. In doing so, we follow the approach employing semantic networks to compute semantic measures also illustrated in [6] and in [65]. In formulae, given two terms w_1 and w_2, each with an associated list of senses $s(w_1)$ and $s(w_2)$, this amounts to computing

$$\operatorname{sim}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\operatorname{sim}(c_1, c_2)] \tag{6}$$

where each conceptual representation must be intended as a vector, as illustrated in Equation (4).

4.2 Experimental Setting and Procedure

Data Sets. As mentioned, the experimentation relies on the MeRaLi system, which has been designed to compute conceptual similarity based on the COVER lexical resource. Its performance has been assessed over four standard data sets. In particular, we considered three data sets for conceptual similarity at the *sense* level,[10] namely the RG [70], MC [52] and WS-Sim data set, which was first designed for conceptual relatedness in [20] and then partially annotated with similarity judgments [1]. Additionally, we considered a fourth dataset recently released in the frame of the SemEval-2017 campaign on Multilingual and Cross-lingual Semantic Word Similarity, and concerned with the computation of the conceptual similarity at the *word* level [7]. Whilst in the former case (sense level conceptual similarity) we computed the similarity by directly applying the formula in Equation (4), in the latter case (word level conceptual similarity) the computation also involves the computation illustrated in Equation (6).

More in detail, the MC data set actually contains 28 pairs, that are a subset of the RG data set, containing 65 sense pairs. The WS-Sim data set is composed of 97 sense pairs, and the Sem-Eval 2017 data set consists of 500 word pairs. The last data set is the most challenging, since it hosts word pairs involving entities. It is challenging also for human common sense in many ways, since it includes pairs such as ⟨Si-o-seh pol, Mathematical Bridge⟩ and ⟨Mount Everest, Chomolungma⟩.

[10] Publicly available at the URL `http://www.seas.upenn.edu/~hansens/conceptSim/`.

Table 3 Percentage of dropped pairs for the *selected data* run of the MeRaLi system.

Dataset	Dropped pairs
MC	17%
RG	15%
WS-Sim	12%
SemEval2017	9%

Evaluation Metrics. The MeRaLi system has then been fed with sense/word pairs, and we recorded the conceptual similarity score provided in output. The similarity scores so obtained have been assessed through Pearson's r and Spearman's ρ correlations, that are usually adopted for the conceptual similarity task. The Pearson r value captures the linear correlation of two variables as their covariance divided by the product of their standard deviations, thus basically allowing to grasp differences in their values, whilst the Spearman ρ correlation is computed as the Pearson correlation between the *rank* values of the considered variables, so it is reputed to be best suited to assess results in a similarity ranking setting where relative scores are relevant [72, 65].

Furthermore, we recorded the output of two runs of the MeRaLi system: in the first one we restricted to considering pairs where the system had enough information on both concepts involved in the comparison (named *selected data* in the following), whilst in the second one we also considered cases where no sufficient information was available in COVER for at least one of the concepts at hand (*full data* in the following). In the former case, we selected the pairs where, for both concepts at stake, a vector description was found in COVER, and at least two shared dimensions were found to be filled (e.g., at least IsA and UsedFor) with at least one element each. Satisfying all these constraints is, in our opinion, necessary in order to be able to justify on which bases two concepts are deemed similar or not. Table 3 shows the percentage of dropped pairs in each data set in the *selected data* condition. Conversely, in the *full data* condition we considered all pairs. In particular, for pairs lacking at least one vector representation, or where less than two shared dimensions were filled, we assigned a similarity score that was set to half the maximum of the evaluation range (that is, in a 0.00-4.00 scale, we set it to 2.00). The rationale underlying these two runs is to try to fully assess the COVER resource, by also investigating to what extent the available information is helpful to conceptual similarity, irrespective of its current coverage, which will be improved in the future releases of the resource.

4.3 Results and Discussion

Table 4 illustrates the results obtained by the MeRaLi system in the experimentation. Compared to the *selected data* run, the strongest competitors in literature obtained 10% higher ρ correlation on the RG data set [65] (3% on the MC data set [1]); 14% on the WS-Sim data set [75]. The distance from

Table 4 Spearman (ρ) and Pearson (r) correlations obtained over the four datasets.

System	RG		MC		WS-Sim		SemEval 2017	
	ρ	r	ρ	r	ρ	r	ρ	r
COVER (Selected data)	0.82	0.88	0.89	**0.91**	0.69	0.70	0.68	0.67
COVER (Full data)	0.76	0.81	0.74	0.79	0.61	0.60	0.65	0.63
$\text{NASARI}_{\text{embed}}$ [7,9,10]	0.88	**0.91**	0.83	**0.91**	0.68	0.68	0.68	0.68
ADW [65]	**0.92**	**0.91**	-	-	0.75	0.72	-	-
PPR [1]	0.83	-	**0.92**	-	-	-	-	-
ConceptNet Numberbatch [75]	-	-	-	-	**0.83**	-	-	-
Luminoso [77]	-	-	-	-	-	-	**0.72**	**0.74**
word2vec [49]	0.84	0.83	-	-	0.78	**0.76**	-	-

state of the art figures is reduced when testing on the SemEval 2017 data set, where we obtained a ρ correlation 4% lower than the Luminoso system [77]. If we consider the *full data* run, our results are some points lower, with minimum (3%) loss w.r.t. the *selected data* run on the SemEval data set.

In order to discuss our results, we focus on the SemEval dataset, that is by far more complete (with 500 word pairs) and varied with respect to the other ones. In fact, it contains named entities and multiword expressions, and covers a wide range of domains.[11]

One major concern is the amount of missing information: as reported in Table 3, almost 10% of word pairs were dropped, as either lacking from COVER or due to the lack of shared information, which prevented us from computing the similarity. Missing concepts may be lacking in (at least one of) the resources upon which the COVER is built: including further resources may thus be helpful to overcome this limitation. Also, integrating further resources in COVER would be beneficial to add further concepts *per* dimension, and to fill more dimensions, so to expand the set of comparisons allowed by the resource.

A discussion of our results on this data set also involves a thorough analysis of the data set itself. The terms in the data set can be naturally arranged into three main classes, involving respectively concept-concept comparisons (400 word pairs), entity-entity comparisons (50 word pairs), and entity-concept pairs (50 word pairs).

So we have re-run the statistical tests to dissect our results according to the three individuated partitions of the data set; the partial results are reported in Table 5.

Entity-Concept pairs. Comparisons involving a concept and an entity are somehow different from those involving only concepts. We individuated two further sub-classes: the pairs where the entity is instance of (that is, in relation InstanceOf with) the class indicated by the concept (e.g., 'Gauss-scientist', 'Harry Potter-wizard', 'NATO-alliance', *etc.*), and cases where the relations

[11] Namely, the 34 domains available in BabelDomains, `http://lcl.uniroma1.it/babeldomains/`.

Table 5 Spearman (ρ) and Pearson (r) correlations (and their harmonic mean) obtained by the MeRaLi system over the three subsets in the *full data* and *selected data* variants.

full data	# pairs	ρ	r	harm.mean
entire data	500	0.65	0.63	0.64
entity-concept	50	0.51	0.45	0.48
entity-entity	50	0.54	0.60	0.57
concept-concept	400	0.67	0.66	0.67
selected data	# pairs	ρ	r	harm.mean
entire data	452	0.68	0.67	0.67
entity-concept	36	0.61	0.60	0.60
entity-entity	31	0.70	0.75	0.72
concept-concept	385	0.68	0.67	0.67

intervening between the two words at stake are not more specific than a general relatedness (e.g., 'Joule-spacecraft', 'Woody Allen-lens', 'islamophobia-ISIS', *etc.*). We then reran the MeRaLi system on the 50 entity-concept pairs (36 pairs in the *selected data* variant), and obtained overall 0.51 ρ correlation (thus significantly lower, than the general figures reported in Table 4). This datum can be complemented by comparing it with the corresponding result in the *selected data* variant: in this case, we obtained 0.61 ρ correlation. Interestingly enough, by focusing on the subset of elements linked by the InstanceOf relationship, we achieved a 0.79 ρ correlation.

These results raise a question. Provided that the InstanceOf relationship is at the base of semantic similarity, COVER is appropriate to unveil semantic similarity for such pairs. However, in the reminder of the entity-concept pairs, the correlation with human judgments is still low. Even more, when the word pairs are not featured by the InstanceOf relationship, it is not simple to understand which sort of comparison is actually being carried out. From a cognitive perspective, it is difficult to follow the strategy adopted by human annotators in providing a similarity score for pairs such as 'Zara-leggings' (gold standard similarity judgement: 1.67 in a 0-4 scale, where 0 is dissimilar and 4 is the identity). In our approach, to assess the similarity between two elements entails individuating under which aspects they can be compared; it means to individuate a set of common properties and relations whose values can be directly compared. This explains that directly comparing a manufacturer and a product is nearly unfeasible, since their features can be hardly compared. In this case it is easy to grasp that the lack of shared (filled) dimensions between the entities may have determined many dropped pairs. Justifying the answer is perhaps helpful to give some information on the argumentative paths that can be possibly followed to assess semantic similarity. One major risk, in these respects, is that instead of *similarity*, the scores provided by human annotators rather refer to generic *relatedness*, which is generally acknowledged as a broader relation than similarity [6]. Similar arguments also apply to meronyms. Let us consider, e.g., the pair 'tail-Boeing 747' (gold standard similarity judgment: 1.92): although each Boeing 747 has a tail, the whole plane (holonym)

cannot be conceptually similar to its tail (meronym), in the same way a car is not similar to one of its wheels.

Entity-Entity pairs. As regards as the entity pairs, in the *selected data* experiment we obtained figures about 15% higher than in the *full data* condition: this is mainly due to the fact that some of the entities were not present in COVER (namely 31 pairs were used in the *selected data* condition vs. the 50 pairs in the *full data* condition). Conversely, the 70% agreement with human annotation is overall a reasonable performance, supporting the appropriateness of COVER. The absence of entities from COVER is easily explained: if either ConceptNet or BabelNet does not contain an element, then this is not present in COVER, that only hosts items that are present in both resources. In order to escape such limitation, next versions of COVER will contain information harvested also from further resources. The rate of agreement obtained experimenting with this subset of data closely approaches — limited to the *selected data* setting — the outstanding results obtained by the Luminoso team at the SemEval 2017 contest [77], and additionally benefits from the explanatory power allowed by the knowledge representation adopted in COVER.

Concept-Concept pairs. This is the principal class in the data set, counting 80% of word pairs in the *full data*, and 96% in the *selected data*. Although also items is this class pose some questions about the concepts at stake (such as comparisons between abstract and concrete entities like the pairs 'coin-payment', 'pencil-story' and 'glacier-global warming'), our results over this subclass of data are by far less sensitive to the filtering performed in the *selected data* experiment (as it is illustrated in Table 5, the results of the MeRaLi system differ about 1% between the two settings). We interpret this result as one corroborating the claim that COVER is mature enough to ensure a reasonable coverage to compute conceptual similarity.

5 Conclusions and Future Work

This article has illustrated COVER, a novel lexical resource, along with COVERAGE, the algorithm designed to build it. COVER puts together the lexicographic precision which is proper to WordNet and BabelNet with the rich common-sense knowledge that features ConceptNet. The obtained vectors capture conceptual information in a compact and cognitively sound fashion. The resource, which basically borrows BabelNet synset IDs as concept identifiers as the naming space, can be easily interfaced to many existing resources that also are linked to BabelNet. We have also shown that COVER is suitable for building NLP applications, in the fields of conceptual categorization, keywords extraction and conceptual similarity. We have reported the results of a thorough experimentation, which was carried out on the conceptual similarity task. Although other approaches presently achieve higher accuracy, the system employing COVER obtains competitive results, and additionally is able

to build explanations of the traits determining the conceptual similarity. The experimentation also revealed that in some cases the information in COVER should be enriched with further information to fully spread the coverage of the resource, and to improve the concept descriptions herein by tuning the balance among the filler dimensions. Another feature that will be added to next releases of the resource is the handling of further languages, thanks to the intrinsically multilingual nature of BabelNet: given that the adopted knowledge in COVER representation is fully conceptual, this step will enable tackling the mentioned tasks in many more languages. Also, the resource to date only contains information on nouns: one fundamental advance will be obtained by accounting for verbs and adjectives whose representation, we believe, will strongly benefit from conceptual information on nouns.

References

1. Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of NAACL*, NAACL '09, pages 19–27. Association for Computational Linguistics, 2009.
2. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735, 2007.
3. Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley Framenet Project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.
4. Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247, 2014.
5. Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63, 2013.
6. Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguists*, 32(1):13–47, 2006.
7. Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, Vancouver, Canada, 2017.
8. José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. A unified multilingual semantic representation of concepts. *Proceedings of ACL, Beijing, China*, 2015.
9. José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. NASARI: a novel approach to a semantically-aware representation of items. In *Proceedings of NAACL*, pages 567–577, 2015.
10. José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.
11. Erik Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2):12–14, 2013.
12. Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10, 2010.
13. Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 168–175. Association for Computational Linguistics, 2003.

14. Davide Colla, Enrico Mensa, and Daniele P. Radicioni. Semantic measures for keywords extraction. In *AI*IA 2017: Advances in Artificial Intelligence*, Lecture Notes for Artificial Intelligence. Springer, 2017.
15. Davide Colla, Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. Tell Me Why: Computational Explanation of Conceptual Similarity Judgments. In *Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Special Session on Advances on Explainable Artificial Intelligence*, Communications in Computer and Information Science (CCIS), Cham, 2018. Springer International Publishing.
16. Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512. IEEE, 2008.
17. Joaquín Derrac and Steven Schockaert. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94, 2015.
18. Ann Devitt and Khurshid Ahmad. Is there a language of sentiment? an analysis of lexical resources for sentiment analysis. *Language resources and evaluation*, 47(2):475–511, 2013.
19. Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.
20. Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001.
21. Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. Multilingual resources for nlp in the lexical markup framework (lmf). *Language Resources and Evaluation*, 43(1):57–70, 2009.
22. Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, 2013.
23. Peter Gärdenfors. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press, 2014.
24. Alexandru-Lucian Gînscă, Emanuela Boroş, Adrian Iftene, Diana TrandabĂţ, Mihai Toader, Marius Corîci, Cenel-Augusto Perez, and Dan Cristea. Sentimatrix: multilingual sentiment analysis service. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 189–195. Association for Computational Linguistics, 2011.
25. Sanda Harabagiu and Dan Moldovan. Question answering. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.
26. Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
27. Catherine Havasi, Robert Speer, and Jason Alonso. ConceptNet: A lexical resource for common sense knowledge. *Recent advances in natural language processing V: selected papers from RANLP*, 309:269, 2007.
28. Eduard Hovy. Text summarization. In *The Oxford Handbook of Computational Linguistics 2nd edition*. Oxford University Press, 2003.
29. Ludovic Jean-Louis, Amal Zouaq, Michel Gagnon, and Faezeh Ensan. An assessment of online semantic annotators for the keyword extraction task. In *Pacific Rim International Conference on Artificial Intelligence*, pages 548–560. Springer, 2014.
30. Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.
31. Sergio Jimenez, Claudia Becerra, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. Softcardinality-core: Improving text overlap with distributional measures for semantic textual similarity. In *Proceedings of *SEM 2013*, volume 1, pages 194–201, 2013.
32. Pat Langley. The cognitive systems paradigm. *Advances in Cognitive Systems*, 1:3–13, 2012.
33. Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

34. Douglas B Lenat, Mayank Prakash, and Mary Shepherd. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine*, 6(4):65, 1985.
35. Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.
36. Antoni Lieto, Andrea Minieri, Alberto Piana, Daniele P. Radicioni, and Marcello Frixione. A dual process architecture for ontology-based systems. In *6th International Conference on Knowledge Engineering and Ontology Development, KEOD 2014*, pages 48–55. INSTICC Press, 2014.
37. Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 48:39–55, 2018.
38. Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 – December 1, 2016, Proceedings*, volume 10037 of *Lecture Notes in Artificial Intelligence*, pages 435–449. Springer, 2016.
39. Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. Taming sense sparsity: a common-sense approach. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.*, 2016.
40. Antonio Lieto, Andrea Minieri, Alberto Piana, and Daniele P. Radicioni. A knowledge-based system for prototypical reasoning. *Connection Science*, 27(2):137–152, 2015.
41. Antonio Lieto and Daniele P. Radicioni. From human to artificial cognition and back: New perspectives on cognitively inspired ai systems. *Cognitive Systems Research*, 39:1–3, 2016.
42. Antonio Lieto, Daniele P. Radicioni, and Valentina Rho. A Common-Sense Conceptual Categorization System Integrating Heterogeneous Proxytypes and the Dual Process of Reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 875–881, Buenos Aires, July 2015. AAAI Press.
43. Antonio Lieto, Daniele P. Radicioni, and Valentina Rho. Dual PECCS: A Cognitive System for Conceptual Representation and Categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):433–452, 2017.
44. Antonio Lieto, Daniele P. Radicioni, Valentina Rho, and Enrico Mensa. Towards a Unifying Framework for Conceptual Represention and Reasoning in Cognitive Systems. *Intelligenza Artificiale*, 11(2):139–153, 2017.
45. Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.
46. Luis Marujo, Ricardo Ribeiro, David Martins de Matos, João P. Neto, Anatole Gershman, and Jaime Carbonell. Key phrase extraction of lightly filtered broadcast news. In *Proceedings of* 15th *International Conference on Text, Speech and Dialogue (TSD 2012)*. Springer, September 2012.
47. John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719, 2012.
48. Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. Merali at semeval-2017 task 2 subtask 1: a cognitively inspired approach. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval 2017)*. Association for Computational Linguistics, 2017.
49. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
50. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
51. George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
52. George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

53. George A Miller and Christiane Fellbaum. Wordnet then and now. *Language Resources and Evaluation*, 41(2):209–214, 2007.
54. David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. In *EMNLP*, pages 262–272. ACL, 2011.
55. Marvin Minsky. Commonsense-based interfaces. *Communications of the ACM*, 43(8):66–73, 2000.
56. Andrea Moro, Francesco Cecconi, and Roberto Navigli. Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 25–28. CEUR-WS. org, 2014.
57. Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.
58. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics, 2010.
59. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
60. David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *The ACM/IEEE Joint Conference on Digital Libraries (JCDL2010)*, Gold Coast, Australia, June 2010. ACM.
61. Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. Different Sense Granularities for Different Applications. In *Proceedings of Workshop on Scalable Natural Language Understanding*, 2004.
62. Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan. Maximizing semantic relatedness to perform word sense disambiguation. *University of Minnesota supercomputing institute research report UMSI*, 25:2005, 2005.
63. Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics, 2004.
64. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
65. Mohammad Taher Pilehvar and Roberto Navigli. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artif. Intell.*, 228:95–128, 2015.
66. Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.
67. Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11(1), 1998.
68. Ray Richardson, Alan F Smeaton, and John Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. In *Proceedings of AICS conference*, pages 1–15, 1994.
69. Eleanor Rosch. Cognitive Representations of Semantic Categories. *Journal of experimental psychology: General*, 104(3):192–233, 1975.
70. Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
71. Hansen A Schwartz and Fernando Gomez. Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In *Procs of the Twelfth Conference on Computational Natural Language Learning*, pages 105–112. ACL, 2008.
72. Hansen A Schwartz and Fernando Gomez. Evaluating semantic metrics on tasks of concept similarity. In *Proc. Int. Florida Artif. Intell. Res. Soc. Conf.(FLAIRS)*, page 324, 2011.
73. Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.
74. Robert Speer and Joshua Chin. An ensemble method to produce high-quality word embeddings. *arXiv preprint arXiv:1604.01692*, 2016.

75. Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451, 2017.
76. Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, pages 3679–3686, 2012.
77. Robert Speer and Joanna Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *CoRR*, abs/1704.03560, 2017.
78. Peter D Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, 2006.
79. Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.
80. Piek Vossen and Christiane Fellbaum. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, chapter Universals and idiosyncrasies in multilingual WordNets. Trends in linguistics / Studies and monographs: Studies and monographs. Mouton de Gruyter, 2009.
81. Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. ACL, 1994.
82. Roman Yampolskiy. Turing test as a defining feature of ai-completeness. *Artificial intelligence, evolutionary computing and metaheuristics*, pages 3–17, 2013.
83. D Yarlett and M Ramscar. Language learning through similarity-based generalization. *Unpublished PhD Thesis, Stanford University*, 2008.