NASARI: a Novel Approach to a Semantically-Aware Representation of Items

José Camacho-Collados, Mohammad Taher Pilehvar and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{collados,pilehvar,navigli}@di.uniroma1.it

Abstract

The semantic representation of individual word senses and concepts is of fundamental importance to several applications in Natural Language Processing. To date, concept modeling techniques have in the main based their representation either on lexicographic resources, such as WordNet, or on encyclopedic resources, such as Wikipedia. We propose a vector representation technique that combines the complementary knowledge of both these types of resource. Thanks to its use of explicit semantics combined with a novel cluster-based dimensionality reduction and an effective weighting scheme, our representation attains state-of-the-art performance on multiple datasets in two standard benchmarks: word similarity and sense clustering. We are releasing our vector representations at http://lcl.uniroma1.it/nasari/.

1 Introduction

Obtaining accurate semantic representations of individual word senses or concepts is vital for several applications in Natural Language Processing (NLP) such as, for example, Word Sense Disambiguation (Navigli, 2009; Navigli, 2012), Entity Linking (Bunescu and Paşca, 2006; Rao et al., 2013), semantic similarity (Budanitsky and Hirst, 2006), Information Extraction (Banko et al., 2007), and resource linking and integration (Pilehvar and Navigli, 2014). One prominent semantic representation approach is the distributional semantic model, which represents lexical items as vectors in a semantic space. The weights in these vectors were traditionally computed on the basis of co-occurrence statistics (Salton et al., 1975; Turney and Pantel, 2010; Dinu and Lapata, 2010; Lappin and Fox, 2014), whereas for the more recent generation of distributional models weight computation is viewed as a context prediction problem, often to be solved by using neural networks (Collobert and Weston, 2008; Turian et al., 2010; Mikolov et al., 2013). Unfortunately, unless they are provided with large amounts of sense-annotated data these corpus-based techniques cannot capture polysemy in their representations, as they conflate different meanings of a word into a single vector. Therefore, most sense modeling techniques tend to base their computation on the knowledge obtained from various lexical resources. However, these techniques mainly utilize the knowledge derived from either WordNet (Banerjee and Pedersen, 2002; Budanitsky and Hirst, 2006; Pilehvar et al., 2013) or Wikipedia (Medelyan et al., 2009; Mihalcea, 2007; Dandala et al., 2013; Gabrilovich and Markovitch, 2007; Strube and Ponzetto, 2006), which are, respectively, the most widely-used lexicographic and encyclopedic resources in lexical semantics (Hovy et al., 2013). This restriction to a single resource brings about two main limitations: (1) the sense modeling does not benefit from the complementary knowledge of different resources, and (2) the obtained representations are resource-specific and cannot be used across settings.

In this paper we put forward a novel concept representation technique, called NASARI, which exploits the knowledge available in both types of resource in order to obtain effective representations of arbitrary concepts. The contributions of this paper are threefold. First, we propose a novel technique

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 567–577,
Denver, Colorado, May 31 – June 5, 2015.

for rich semantic representation of arbitrary WordNet synsets or Wikipedia pages. Second, we provide improvements over the conventional *tf-idf* weighting scheme by applying lexical specificity (Lafon, 1980), a statistical measure mainly used for term extraction, to the task of computing vector weights in a vector representation. Third, we propose a semantically-aware dimensionality reduction technique that transforms a lexical item's representation from a semantic space of words to one of WordNet synsets, simultaneously providing an implicit disambiguation and a distribution smoothing. We demonstrate that our representation achieves state-of-the-art performance on two different tasks: (1) word similarity on multiple standard datasets: MC-30, RG-65, and WordSim-353 similarity, and (2) Wikipedia sense clustering, in which our unsupervised system surpasses the performance of a state-of-the-art supervised technique that exploits knowledge available in Wikipedia in several languages.

2 Semantic Representation of Concepts

Lexical resources and concepts. The gist of our approach lies in its combination of knowledge from two different lexical resources: (1) the expert-based lexicographic WordNet, whose basic constituents are synsets, i.e., concepts expressed by sets of synonymous words (Miller et al., 1990), and (2) the collaboratively-constructed encyclopedic Wikipedia, whose articles can be considered as individual concepts. Throughout the paper, by a concept we mean a tuple $b = (p, s)$ where p is a Wikipedia page and s is the corresponding WordNet synset. As a bridge between the two resources we use the synset-to-article mappings provided by BabelNet[1] (Navigli and Ponzetto, 2012), a high coverage multilingual encyclopedic dictionary and semantic network that merges, among other resources, Wikipedia and WordNet. Note that the concept b can also contain a Wikipedia page or a WordNet synset only, if a mapping is not provided by BabelNet.

Semantic representation: NASARI. Our concept modeling approach consists of two phases. First, for a given concept, we collect a set of relevant Wikipedia pages by leveraging the structural information in Wikipedia and WordNet (Section 2.1).

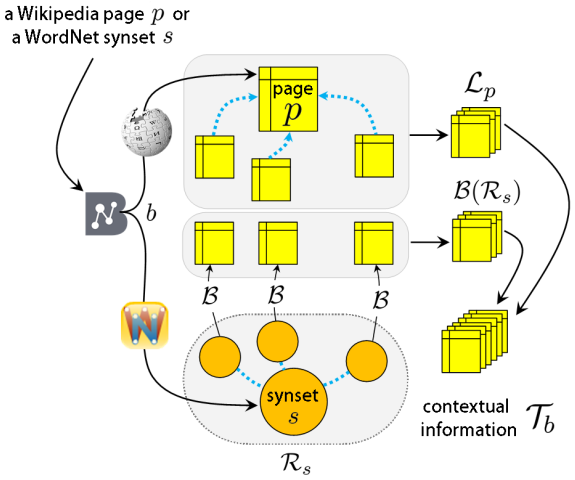

Figure 1: The process of obtaining contextual information for a WordNet synset or a Wikipedia article.

Then, we analyze the obtained contextual information and construct two vector representations of the concept (Section 2.2).

2.1 Collecting contextual information

Figure 1 illustrates the process of obtaining a set of relevant Wikipedia pages $\mathcal{T}_b$ as contextual information for a given concept $b = (p, s)$. Let $\mathcal{L}_p$ be the set containing p and all the Wikipedia pages having an outgoing link to p, and $\mathcal{R}_s$ be the set consisting of s and all other synsets that are in its direct neighbourhood. We further enrich $\mathcal{R}_s$ by including the coordinate synsets of s and the related synsets from its disambiguated gloss[2]. Let $\mathcal{B}$ be a function mapping each WordNet synset s' to its corresponding Wikipedia page p, if such mapping exists in BabelNet, and to the empty set otherwise. Hence, $\mathcal{B}(\mathcal{R}_s) = \cup_{s' \in \mathcal{R}_s} \mathcal{B}(s')$. Then, our contextual information is the set of Wikipedia pages $\mathcal{T}_b = \mathcal{L}_p \cup \mathcal{B}(\mathcal{R}_s)$. In the case either p or s is not present in the concept b, we take the contextual information as $\mathcal{T}_b = \mathcal{B}(\mathcal{R}_s)$ or $\mathcal{T}_b = \mathcal{L}_p$, respectively.

2.2 Vector construction

By processing the collected contextual information $\mathcal{T}_b$, NASARI represents the concept b as two vectors in two semantic spaces: (1) word-based and (2) synset-based. Let $\mathcal{W}_b$ be the bag of words of all the Wikipedia pages in $\mathcal{T}_b$ after lemmatization and stop-

[1] http://www.babelnet.org/

[2] http://wordnet.princeton.edu/glosstag.shtml

word removal. We use lexical specificity in order to extract the most representative words (Section 2.2.1) and synsets (Section 2.2.2) of $\mathcal{W}_b$.

Lexical specificity. Lexical specificity (Lafon, 1980) is a statistical measure that has been used in a wide range of NLP applications, such as textual data analysis (Lebart et al., 1998), term extraction (Drouin, 2003), and domain disambiguation (Camacho Collados et al., 2014). However, to our knowledge, it has never heretofore been used to calculate weights in a vector-based representation (Turney and Pantel, 2010). Lexical specificity is based on the hypergeometric distribution over word frequencies. This statistical measure is particularly suitable for extracting an accurate set of representative terms for a given subcorpus of a reference corpus (Lafon, 1980). Unlike the conventional term frequency-inverse document frequency weighting scheme (Jones, 1972, *tf-idf*), lexical specificity is not especially sensitive to different text lengths.

Assume a reference corpus of T words and a t-words subcorpus of that corpus. The goal is to find a set of terms that are peculiar to the subcorpus, but not to the whole reference corpus. Formally, given a word w that occurs f and k times in the corpus and subcorpus, respectively, positive specificity computes the relevance of w to the subcorpus as $P(X \geq k)$ if $k \geq \frac{ft}{T}$, where X is a random variable following a hypergeometric distribution with parameters f, t and T, and $\frac{ft}{T}$ is the expected value of X. In our setting we are only interested in the positive specificity, i.e., the set of most relevant words appearing in the contextual information. We apply the standard procedure of applying log_{10} and then inverting the sign of the specificity probabilities in order to re-scale them to the real line, which is more easily interpretable (Drouin, 2003; Camacho Collados et al., 2014). We only retain words with specificity greater than two, which is equal to $-log_{10}(0.01)$. This threshold leads to a set of representative words that are relevant to the context with a confidence of at least 99%, i.e., $P(X \geq k) \leq 0.01$ (Billami et al., 2014).

2.2.1 Word-based representation

This word-based representation models the concept b in a conventional semantic space whose dimensions are individual words. We leverage lexical specificity to compute a weighted set of most representative words for $\mathcal{W}_b$ with respect to the reference corpus, i.e., the whole Wikipedia. As an example, the obtained word-based vector for the *edge of water* sense of *shore* has *water*, *ocean*, *lake*, *beach* and *sea* among its most relevant dimensions.

2.2.2 Synset-based representation

Given that the amount of contextual information gathered for a concept can be small, the resulting word-based vector can be sparse and as a consequence prone to noise, especially in the case of less frequent concepts. Therefore, we put forward a method that tackles the issue, providing rich semantically-aware representations. To this end, we group - and thereby generalize - similar dimensions in the obtained word-based vector, to produce a smaller vector in which dimensions are WordNet synsets and weights are computed on the basis of the combined information of the individual words in the group. The generalization procedure can be summarized in two main steps.

First, for each word w in $\mathcal{W}_b$, we obtain from WordNet the set $\mathcal{H}_w$ of all the direct hypernyms of all the synsets containing w. For each synset $h \in \mathcal{H}_w$ we check whether there exists another word w' from the contextual information that is a hyponym of h, i.e., $h \in \mathcal{H}_w \cap \mathcal{H}_{w'}$. When such is the case, letting $\mathcal{Y}_h$ be the set of all words in the hyponym synsets of h, we combine w, w' and all the other words in $\mathcal{Y}_h$ into a single dimension represented by their common hypernym h. Thus for our earlier example, the three words *ocean*, *lake*, and *sea* are grouped into a single dimension represented by their hypernym, i.e., the synset containing the *body of water* sense of *water* (*water*^2_n in WordNet 3.0)[3].

Then, we compute the weight associated with the new dimension by calculating the lexical specificity of the word cluster. Formally, we calculate the lexical specificity of h by setting the parameters k and f as the total number of times the words in $\mathcal{Y}_h$ occur in $\mathcal{W}_b$ and the whole Wikipedia, respectively. The values of t and T remain unchanged.

Our generalization procedure is similar to the dimensionality reduction that is performed using singular value decomposition in Latent Semantic Analysis (Landauer and Dumais, 1997, LSA). However,

[3]We denote the i^{th} sense of word w with POS p as w^i_p.

LSA is not applicable to our setting because, due to the usage of lexical specificity, our vectors are relatively small in size and different vectors usually have few overlapping dimensions. Moreover, unlike LSA, in which the size-reduced vectors have opaque conflations of multiple words as their dimensions, our new semantic space has human- and machine-readable synsets as its dimensions. Our generalization procedure produces three advantages: (1) it maps the vectors from a word-based semantic space to a lower-dimensional space of synsets, (2) while merging multiple words into a single synset an implicit disambiguation of context words takes place, providing better means for sense distinction, and (3) the dimensionality reduction tackles the potential noise and sparsity, resulting in smoother vectors.

3 NASARI for Semantic Similarity

So far we have explained how NASARI constructs two types of representations, i.e., word-based and synset-based, for arbitrary WordNet synsets and Wikipedia pages. In this section we provide a method that leverages NASARI representations for effective measurement of concept and word similarity. Semantic similarity between a pair of lexical items (e.g., words or concepts) lies at the core of many applications in NLP and hence it has received a considerable amount of research interest, leading to a wide range of semantic similarity measures (Mohammad and Hirst, 2012).

3.1 Concept similarity

Given a pair of concepts, we first use the procedure described in Section 2 to obtain for each concept the two corresponding vector representations, i.e., word-based and synset-based. For each representation type, we then compute the similarity of the two concepts by comparing their corresponding vectors. This results in two similarity scores, one for each representation type. The final similarity is computed as the average of the two similarity scores. We use Weighted Overlap for comparing vectors.

Weighted Overlap. Proposed by Pilehvar et al. (2013), Weighted Overlap (WO) first sorts the elements of each vector v_i and then harmonically weights the overlaps between the two vectors:

$$WO(v_1, v_2) = \frac{\sum_{q \in O}(r_q^1 + r_q^2)^{-1}}{\sum_{i=1}^{|O|}(2i)^{-1}} \qquad (1)$$

where O is the set of overlapping dimensions between the two vectors and r_q^j is the rank of dimension q in the vector v_j. Given that our vectors are significantly smaller than those in the original setting of WO, the overlaps are also generally smaller in size. Hence, we apply a square root operation to the computed value in order to obtain a more uniformly-distributed range of scores across the similarity scale, i.e., [0, 1]. In our experiments we show the advantage we gain by using WO in comparison to the conventional cosine measure.

Algorithm 1 NASARI-based word similarity

Input: words w_1 and w_2
Output: *Sim*, similarity score

for each synonym set $H \in \mathcal{S}$
  if $w_1 \in H$ & $w_2 \in H$ **then**
    return $Sim = 1$
for each word $w_i \in \{w_1, w_2\}$
  $\mathcal{C}_{w_i} \leftarrow \emptyset$, set of concepts associated with w_i
  if $w_i \in$ *WordNet* & w_i *not* Named Entity **then**
    for each sense $s \in$ *WordNet senses* (w_i)
      $\mathcal{C}_{w_i} \leftarrow \mathcal{C}_{w_i} \cup \{s\}$
  else
    for each page $p \in$ *piped links* (w_i)
      $\mathcal{C}_{w_i} \leftarrow \mathcal{C}_{w_i} \cup \{p\}$
  $V_i \leftarrow \emptyset$, set of representations for concepts in $\mathcal{C}_{w_i}$
  for each concept $c \in \mathcal{C}_{w_i}$
    $v_{wrd} \leftarrow$NASARI word-based rep. of c
    $v_{syn} \leftarrow$NASARI synset-based rep. of c
    $v \leftarrow (v_{wrd}, v_{syn})$
    $V_i \leftarrow V_i \cup \{v\}$
$Sim \leftarrow \max_{v \in V_1, v' \in V_2} \frac{WO(v_{wrd}, v'_{wrd}) + WO(v_{syn}, v'_{syn})}{2}$
return *Sim*

3.2 Word similarity

Algorithm 1 shows the procedure we devised for measuring semantic similarity between two words. There are three main steps:

1. Given a pair of words w_1 and w_2 the algorithm first checks whether they are synonymous according to our synonym set collection $\mathcal{S}$. In Section 3.2.1, we explain how we obtain this set. If the words are defined as synonyms in $\mathcal{S}$,

the algorithm returns the maximum similarity score of one (lines 1-3).

2. If the words are not defined as synonyms, we proceed by obtaining, for each word w_i, its set of possible senses ($\mathcal{C}_{w_i}$, lines 5-11). We accordingly obtain the set of their respective NASARI vector representations (V_i, lines 13-17), two (word-based and synset-based) for each concept in $\mathcal{C}_{w_i}$. Section 3.2.2 describes the concept extraction process.

3. Finally, the algorithm returns the similarity score *Sim* (line 19), calculated as the similarity of the closest senses of w_1 and w_2. In our default setting, we linearly combine our two vector representations by averaging them (line 18).

3.2.1 Wiktionary synonyms $\mathcal{S}$

Wiktionary is a rich collaboratively-constructed lexical resource that provides a considerable amount of multilingual lexical information for a large number of words. We use this resource in order to obtain sets of synonymous words $\mathcal{S}$. To this end, we first extract all the pre-specified synonymy relations in the English Wiktionary. This results in 17K sets with an average size of 2.8 synonyms.

In order to enrich the set we introduce a method that exploits the multilinguality of Wiktionary to extract synonymous words. Our approach utilizes translations of words in other languages as bridges between synonymous words in English. Specifically, for each sense s of word w in Wiktionary, we first get all the available translations. Assume that the sense s of w translates into the word t_l in language l. If there is another word sense s' of another word w' in Wiktionary that is also translated to t_l in language l, we hypothesize that w and w' are synonyms. In order to avoid ambiguity, as t_l we only consider words that are monosemous according to language l.

This procedure results in around 9K additional synonymous sets with an average size of 2.1. For instance, the Finnish noun *ammatti*, which is monosemous according to Wiktionary, links seven English words into a single set of synonyms: *career*, *business*, *profession*, *occupation*, *trade*, *calling*, and *vocation*. The final synonym set collection $\mathcal{S}$ contains 25K sets, each having, on average, 2.6 words.

3.2.2 Concept extraction

If the two input words w_1 and w_2 are not found in the same synonym set in $\mathcal{S}$, we proceed by obtaining their sets of senses $\mathcal{C}_{w_1}$ and $\mathcal{C}_{w_2}$, respectively. Depending on the type of w_i, we use two different resources for obtaining $\mathcal{C}_{w_i}$: the WordNet sense inventory and Wikipedia.

WordNet words. When the word w_i is defined in the WordNet sense inventory and is not a named entity (line 6 in Algorithm 1), we set $\mathcal{C}_{w_i}$ as all the WordNet synsets that contain w_i, i.e., $\mathcal{C}_{w_i} = \{\textit{synset } s \in \textit{WordNet} : w_i \in s\}$. We use Stanford Named Entity Recognizer (Finkel et al., 2005) in our experiments.

WordNet OOV and named entities. For named entities and words that do not exist in WordNet's vocabulary (OOV) we construct the set $\mathcal{C}_{w_i}$ by exploiting Wikipedia's piped links (line 10 in Algorithm 1). To this end, we take as elements of $\mathcal{C}_{w_i}$ the Wikipedia pages of the hyperlinks which have w_i as their surface form, i.e., *piped-links* (w_i). If $|\mathcal{C}_{w_i}| > 5$, we prune $\mathcal{C}_{w_i}$ to its top-5 pages in terms of their number of ingoing links. Our choice of Wikipedia as a source for named entities is due to its higher coverage in comparison to WordNet.

4 Experiments

We evaluated NASARI on two different tasks that require the computation of semantic similarity between words or concepts: word similarity (Section 4.1) and sense clustering (Section 4.2).

4.1 Word similarity

4.1.1 Datasets

We took as benchmark for our word similarity experiments three standard datasets that are widely used in the literature: RG-65 (Rubenstein and Goodenough, 1965), MC-30 (Miller and Charles, 1991), and WordSim-353 (Finkelstein et al., 2002; Agirre et al., 2009). WordSim-353 originally conflated similarity and relatedness, leading to high similarity scores for pairs such as *computer-keyboard* despite the dissimilarity in their meanings. To correct the conflation, Agirre et al. (2009) partitioned the dataset into two subsets: relatedness and similarity. Given that our similarity measure is targeted at semantic similarity, we took the similarity subset of

WordSim-353 (WS-Sim) as test bed for our evaluations. The subset comprises 203 word pairs.

4.1.2 Experimental setup

In this task, we assess the performance of different systems in terms of Pearson correlation. We compare our system against six similarity measures that have reported best performance on the three datasets. Lin (Lin, 1998) and ADW (Pilehvar et al., 2013) are WordNet-based approaches that leverage the structural information of WordNet for the computation of semantic similarity. Most similar to our work are Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007, ESA), which represents a word in a high-dimensional space of Wikipedia articles, and Salient Semantic Analysis (Hassan and Mihalcea, 2011, SSA), which leverages the linking of concepts within Wikipedia articles for generating *semantic profiles* of words. Word2Vec (Mikolov et al., 2013) and PMI-SVD are the best predictive and co-occurrence models obtained by Baroni et al. (2014) on a 2.8 billion-token corpus that also includes the English Wikipedia.[4] Word2Vec is based on neural network context prediction models (Mikolov et al., 2013), whereas PMI-SVD is a traditional co-occurrence based vector wherein weights are calculated by means of Pointwise Mutual Information (PMI) and the vector's dimension is reduced to 500 by singular value decomposition (SVD). We use the DKProSimilarity (Bär et al., 2013) implementation of Lin and ESA in order to evaluate these measures on the WS-Sim dataset.

4.1.3 Results

Table 1 shows the Pearson correlation of the different similarity measures on the three datasets considered. NASARI proves to be highly reliable on the task of word similarity, providing state-of-the-art performance on RG-65 and MC-30, and competitive results on WS-Sim. Importantly, the improvement we attain over measures that utilize as their knowledge base either WordNet (i.e., ADW, Lin) or Wikipedia (i.e., ESA and SSA) shows that our usage of the complementary information of the two types of resource has been helpful. We note that our Wiktionary module detects four additional synonymous pairs (i.e., similarity = 1.0) in MC-30 (13%), eight in RG-65 (12%), and thirteen in WS-Sim (6%) that are not defined as synonyms in WordNet. We also obtain competitive results according to the Spearman correlation (a setting in which the absolute similarity scores do not play a role and it is solely their ranking that matters) on all the three datasets: MC-30 (0.89), RG-65 (0.88), and WS-Sim (0.73).

Measure	RG-65	MC-30	WS-Sim
NASARI	**0.91**	**0.91**	0.74
SSA	0.86	0.88	NA
Word2Vec	0.84$^{\diamond}$	0.83‡	**0.76**‡
Lin	0.83	0.82	0.66
ADW	0.81	0.79	0.63
PMI-SVD	0.74$^{\diamond}$	0.76‡	0.68‡
ESA	0.72	0.74	0.45

Table 1: Pearson correlation of different similarity measures on RG-65, MC-30, and WordSim-353 similarity (WS-Sim) datasets. Results for Lin and ESA on RG-65 and MC-30 are taken from (Hassan and Mihalcea, 2011). We show the best performance obtained by Baroni et al. (2014) out of 48 configurations specifically tested on RG-65 (highlighted by $\diamond$) and across different datasets including WS-Sim (highlighted by $\ddagger$).

WS-Sim is the only dataset on which we do not report state-of-the-art performance. An analysis of the outputs of our system on the WS-Sim dataset revealed that there are pairs in this subset of WordSim-353 that are not assigned proper scores according to the similarity scale. Hill et al. (2014) had previously pointed out this deficiency of WS-Sim, mainly due to its original relatedness-based scoring scale. For instance, word pairs that are barely related (e.g., *street-children*) or antonyms (e.g., *profit-loss* and *smart-stupid*) are assigned relatively high similarity values (respectively, 4.9 for the former and 7.3 and 5.8 for the latter case, in the 0-10 scale). In all these cases our system produces more appropriate judgements according to the similarity scale. On the other hand, there are highly similar pairs in the dataset with relatively low gold scores. Examples include *school-center*[5] and *term-life*[6] with the respective gold similarity scores of 3.4 and 4.5, whereas

[4] clic.cimec.unitn.it/composes/semantic-vectors.html

[5] *School* and *center* have a pair of highly similar senses in WordNet 3.0: *center*$^{3}_{n}$*: "a building dedicated to a particular activity"* and *school*$^{2}_{n}$*: "a building where young people receive education."*

[6] *Term* and *life* are in coordinate synsets (with *time period* as their common hypernym) in WordNet 3.0.

Measure	System type	500-pair	SemEval
NASARI	unsupervised	**84.6%**	**88.4%**
Dan-mono	supervised	77.4%	83.5%
Dan-multi	supervised	84.4%	85.5%
Baseline	-	71.4%	82.5%

Table 2: Accuracy of different systems on two manually-annotated English datasets for sense clustering in Wikipedia. Dan-mono and Dan-multi are the monolingual and multilingual systems of Dandala et al. (2013).

NASARI computes their similarities as 8.4 and 9.6.

4.2 Sense clustering

Our second set of experiments focuses on sense clustering of the Wikipedia sense inventory. Wikipedia can be considered as a sense inventory wherein the different meanings of a word are denoted by the articles listed in its disambiguation page (Mihalcea and Csomai, 2007). Given the high granularity of this inventory, clustering of senses can be highly beneficial to tasks that take this encyclopedic resource as their sense inventory (Hovy et al., 2013), such as Wikipedia-based Word Sense Disambiguation (Mihalcea, 2007; Dandala et al., 2013).

4.2.1 Datasets

For the sense clustering task, we take as our benchmark the two datasets created by Dandala et al. (2013). In these datasets, clustering has been viewed as a binary classification problem in which all possible pairings of senses of a word are annotated whether they ought to be clustered or not. The first dataset contains 500 pairs, 357 of which are set to *clustered* and the remaining 143 to *not clustered*. The second dataset, referred to as the SemEval dataset, is based on a set of highly ambiguous words taken from SemEval evaluations (Mihalcea, 2007) and consists of 925 pairs, 162 of which are positively labeled, i.e., clustered.

4.2.2 Experimental setup

In this task we use the procedure explained in Section 3.1 for measuring the similarity of concepts. A pair of pages is set to belong to the same cluster if their similarity exceeds the middle point in our similarity scale, i.e., 0.5 in the scale of [0, 1]. We compare our results with the state-of-the-art systems of Dandala et al. (2013) that perform clustering by exploiting the structure and content of an English page (monolingual variant), or several pages in different languages (multilingual variant that uses English, German, Spanish and Italian pages). These systems are essentially multi-feature Support Vector Machine classifiers that use an automatically-labeled dataset for their training.

4.2.3 Results

Table 2 lists the results of NASARI as well as the state-of-the-art systems of Dandala et al. (2013). We also report the results for a baseline system that sets all pairs as *not clustered*. As can be seen from the table, our system proves to be highly robust and competitive by outperforming, in an unsupervised setting, the supervised monolingual and multilingual systems of Dandala et al. (2013) on both datasets. As regards the F1, we obtain 72.0% and 64.2% on the 500-pair and SemEval datasets, respectively, a measure that is not reported by Dandala et al. (2013).

4.3 Analysis

Recall from Section 2 that our system has two vector representations, for each of which we compute vectors based on lexical specificity. We also mentioned in Section 3 that we opt for Weighted Overlap as our vector comparison method. In order to analyze the impact of each of these elements, we carried out a series of experiments with the conventional logarithmically-scaled *tf-idf* weighting scheme and the cosine vector comparison technique. For a word w, we calculate the *tf-idf* by taking *tf* as the frequency of w in the corresponding contextual information, and $idf = log(|D|/|\{p \in D : w \in p\}|)$, where D is the set of all pages in Wikipedia.

Table 3 shows the performance of the NASARI-based similarity system and its individual vector representations for different weight computation schemes, i.e., lexical specificity and *tf-idf*, and for different vector comparison techniques, i.e., cosine and WO, on word similarity and sense clustering datasets. As can be seen from the Table, the performance of the word-based representation consistently improves on both tasks when combined with the additional information from the synset-based vectors, demonstrating that the sense distinctions offered by the generalization process have been beneficial.

Between the two vector comparison methods, WO proves to better suit our specificity-based vec-

Vector representation	Weighting scheme	Vector comparison	Word similarity			Sense clustering	
			MC-30	RG-65	WS-Sim	500-pair	SemEval
Combined	specificity	WO	⋆**0.91**	⋆**0.91**	⋆0.74	⋆**84.6%**	⋆**88.4%**
		cosine	0.88	0.89	**0.75**	76.2%	83.6%
	tf-idf	WO	0.85	0.87	0.73	60.4%	67.8%
		cosine	0.79	0.84	0.70	81.4%	86.1%
Word-based	specificity	WO	**0.90**	**0.91**	**0.73**	**82.0%**	**85.0%**
		cosine	0.86	0.88	0.72	73.2%	83.4%
	tf-idf	WO	0.86	0.87	0.72	78.4%	82.6%
		cosine	0.83	0.87	0.71	79.2%	84.4%
Synset-based	specificity	WO	**0.91**	**0.90**	**0.75**	78.8%	83.8%
		cosine	0.90	0.88	**0.75**	**79.8%**	**85.0%**
	tf-idf	WO	0.86	0.85	0.73	37.2%	41.1%
		cosine	0.71	0.80	0.66	79.4%	**85.0%**
Word-based	specificity	WO	†0.86	†0.87	†0.71	†80.0%	†85.1%

Table 3: Performance of NASARI and its individual vector representations for different weight computation schemes, i.e., lexical specificity and *tf-idf*, and for different vector comparison techniques, i.e., cosine and Weighted Overlap (WO), in terms of Pearson correlation (word similarity) and accuracy (sense clustering). The scores highlighted by ⋆ are the ones obtained using our default NASARI setting, and the ones highlighted by † correspond to the setting of our system using Wikipedia as its only knowledge source.

tors by outperforming cosine in most cases. The reason behind the lower performance of WO for the synset-based vectors on the task of sense clustering can be explained by the nature of the corresponding datasets. Since the synset-based vectors and their overlapping dimensions are small, their cosine similarity scores also tend to be relatively low, unlike WO whose range of values is not affected by the number of overlapping dimensions. Given that in the experiments the threshold is fixed to the middle point of the scale (cf. Section 4.2.2), generally low similarity values lead to a high-precision, low-recall system, which is rewarded by higher accuracy performance in datasets in which a large portion of instances are negative. In fact, for the synset-based vector representation weighted using specificity, the F1 performance of the cosine is significantly lower than WO. On the SemEval dataset the F1 performance of WO is 60.1%, whereas cosine attains 37.1%. Similarly, on the 500-pair dataset, WO leads cosine by 16.8%: 68.5% vs. 51.7%.

As far as the weighting scheme is concerned, lexical specificity outperforms *tf-idf* on both tasks, irrespective of the vector comparison technique and representation. We attribute the better performance of lexical specificity to the probabilistic nature of weights in its vectors. The *tf-idf* weighting scheme, in contrast, suffers from insensitivity to the relative size of the contextual information. Thus, subsequently, specificity-based vectors provide the advantage of accurately reducing the vectors' dimension, unlike the *tf-idf* scheme in which the size-insensitive weights are not comparable across vectors. As a result, the specificity-based vectors are substantially smaller in size, bringing about better space utilization and faster running time. In our experiments the vectors obtained by using lexical specificity were, on average, almost nine times (2505 vs. 21825) and four times (335 vs. 1311) smaller than the *tf-idf*-based vectors for the word-based and synset-based vector representations, respectively.

We were also interested in verifying the advantage gained by combining the complementary knowledge of Wikipedia and WordNet. To this end, we carried out an experiment in which NASARI uses Wikipedia as its only knowledge source (i.e., without using WordNet). The last row in the Table (highlighted by †) shows the results for this setting. Note that since WordNet is not used in this setting, we are constrained to the word-based vector representation only. The results show that the combination of the types of resource leads to a consistent performance

improvement across tasks and datasets, with the average improvement being 5%.

5 Related Work

Given that in this work we focused mainly on similarity for the evaluation of our semantic representation, in addition to concept representation, we also briefly discuss related works for semantic similarity.

Concept representation. Distributional semantic models are usually the first choice for representing textual items such as words or sentences (Turney and Pantel, 2010). These models have attracted considerable research interest, resulting in various co-occurrence based representations (Salton et al., 1975; Evert, 2005; Pado and Lapata, 2007; Erk and Padó, 2008) or predictive models (Collobert and Weston, 2008; Turian et al., 2010; Mikolov et al., 2013; Baroni et al., 2014). Although there have been approaches proposed in the literature for learning sense-specific embeddings (Weston et al., 2013; Huang et al., 2012; Neelakantan et al., 2014), their coverage is limited only to those senses that are covered in the underlying corpus. Moreover, the obtained sense representations are usually not linked to any sense inventory, and therefore such linking has to be carried out, either manually, or with the help of sense-annotated data. Hence, unless they are provided with large amounts of sense-annotated data, these techniques cannot furnish an effective representation of word senses in an existing standard sense inventory.

Consequently, most sense modeling techniques have based their representation on the knowledge derived from resources such as WordNet (Mihalcea and Moldovan, 1999; Agirre and Lopez, 2003; Agirre and de Lacalle, 2004; Pilehvar et al., 2013), or Wikipedia (Gabrilovich and Markovitch, 2007; Mihalcea, 2007). None of these techniques, however, combine knowledge from multiple types of resource, making their representations resource-specific and also prone to sparsity. In contrast, our method is based on the complementary knowledge of two different resources and their interlinking, leading to richer semantic representations that are also applicable across resources. Most similar to our combination of complementary knowledge is the work of Franco-Salvador et al. (2014) for cross-lingual document retrieval.

Concept similarity. Concept similarity techniques are mainly limited to the knowledge that their underlying lexical resources provide. For instance, methods designed for measuring semantic similarity of WordNet synsets (Banerjee and Pedersen, 2002; Budanitsky and Hirst, 2006; Pilehvar et al., 2013) usually leverage lexicographic or structural information in this lexical resource. Similarly, Wikipedia-based approaches (Hassan and Mihalcea, 2011; Strube and Ponzetto, 2006; Milne and Witten, 2008) do not usually benefit from the expert-based lexico-semantic knowledge provided in WordNet. In contrast, our approach combines knowledge from both resources, providing two advantages: (1) more effective measurement of similarity based on rich semantic representations, and (2) the possibility of measuring cross-resource semantic similarity, i.e., between Wikipedia pages and WordNet synsets.

6 Conclusions

In this paper we presented a novel semantic approach, called NASARI, for effective vector representation of arbitrary WordNet synsets and Wikipedia pages. The strength of our approach lies in its combination of complementary knowledge from different types of resource, while at the same time it also benefits from an effective vector representation with two novel features: lexical specificity for the calculation of vector weights and a semantically-aware dimensionality reduction. NASARI attains state-of-the-art performance on multiple standard benchmarks in word similarity as well as Wikipedia sense clustering. We release the representations obtained for all the Wikipedia pages and WordNet synsets in http://lcl.uniroma1.it/nasari/. As future work we plan to integrate NASARI into BabelNet and apply our representation to a multilingual setting, enabling the comparison of pairs of concepts across languages. We also intend to use our approach on the task of multilingual Word Sense Disambiguation.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234.

References

Eneko Agirre and Oier Lopez de Lacalle. 2004. Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of LREC*, pages 1123–1126, Lisbon, Portugal.

Eneko Agirre and Oier Lopez. 2003. Clustering WordNet word senses. In *Proceedings of RANLP*, pages 121–130, Borovets, Bulgaria.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27, Boulder, Colorado.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for Word Sense Disambiguation using WordNet. In *Proceedings of CICLing*, pages 136–145, Mexico City, Mexico.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the Web. In *Proceedings of IJCAI*, pages 2670–2676, New York, NY, USA.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. DKPro similarity: An open source framework for text similarity. In *Proceedings of ACL: System Demonstrations*, pages 121–126, Sofia, Bulgaria, August.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, Maryland.

Mokhtar-Boumeyden Billami, José Camacho-Collados, Evelyne Jacquey, and Laurence Kister. 2014. Semantic annotation and terminology validation in full scientific articles in social sciences and humanities (annotation sémantique et validation terminologique en texte intégral en shs) [in french]. In *Proceedings of TALN 2014*, pages 363–376, Marseille, France.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL*, pages 9–16, Trento, Italy.

Jose Camacho Collados, Mokhtar Billami, Evelyne Jacquey, and Laurence Kister. 2014. Approche statistique pour le filtrage terminologique des occurrences de candidats termes en texte intgral. In *Journées internationales d'Analyse statistique des Données Textuelles*, pages 121–133, Paris, France.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167, Helsinki, Finland.

Bharath Dandala, Chris Hokamp, Rada Mihalcea, and Razvan C. Bunescu. 2013. Sense clustering using Wikipedia. In *Proceedings of RANLP*, pages 164–171, Hissar, Bulgaria.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of EMNLP*, pages 1162–1172, Massachusetts, USA.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, pages 897–906, Edinburgh, UK.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, Universitt Stuttgart.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370, Ann Arbor, Michigan.

Lev Finkelstein, Gabrilovich Evgeniy, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the* 14^{th} *Conference on European chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, Hyderabad, India.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *Proceedings of AAAI*, pages 884,889, San Francisco, USA.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. arXiv:1408.3456.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882, Jeju Island, Korea.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1:127–165.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Shalom Lappin and Chris Fox, editors. 2014. *Handbook of Contemporary Semantics – second edition*. Wiley-Blackwell, Malden, MA.

Ludovic Lebart, Andre Salem, and Lisette Berry. 1998. *Exploring textual data*. Kluwer Academic Publishers.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of ICML*, pages 296–304, San Francisco, CA.

Olena Medelyan, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.

Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of ACM on Information and Knowledge management*, pages 233–242, Lisbon, Portugal.

Rada Mihalcea and Dan Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI*, pages 461–466, Orlando, Florida, USA.

Rada Mihalcea. 2007. Using Wikipedia for automatic Word Sense Disambiguation. In *Proceedings of NAACL-HLT-07*, pages 196–203, Rochester, NY.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at International Conference on Learning Representations*, Scottsdale, Arizona.

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

George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4):235–244.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08*, pages 25–30, Chicago, IL.

Saif Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR*, abs/1203.1858.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Roberto Navigli. 2012. A quick tour of Word Sense Disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pages 1059–1069, Doha, Qatar.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of ACL*, pages 468–478, Baltimore, USA.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351, Sofia, Bulgaria.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Gerard Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1419–1424, Boston, Massachusetts.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, pages 384–394, Uppsala, Sweden.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In *Proceedings of EMNLP*, pages 1366–1371, Seattle, Washington, USA.