Natural Language Engineering **1** (1): 1–30. Printed in the United Kingdom

Similarity Computation Using Semantic Networks Created From Web-Harvested Data

ELIAS IOSIF and ALEXANDROS POTAMIANOS
Department of Electronic and Computer Engineering, Technical University of Crete
Chania 73100, Greece
email: `{iosife,potam}@telecom.tuc.gr`

(*Received 28 February 2013*)

Abstract

We investigate language-agnostic algorithms for the construction of unsupervised distributional semantic models using web-harvested corpora. Specifically, a corpus is created from web document snippets and the relevant semantic similarity statistics are encoded in a semantic network. We propose the notion of semantic neighborhoods that are defined using co-occurrence or context similarity features. Three neighborhood-based similarity metrics are proposed, motivated by the hypotheses of attributional and maximum sense similarity. The proposed metrics are evaluated against human similarity ratings achieving state-of-the-art results.

1 Introduction

Semantic similarity is the building block for numerous applications of natural language processing (NLP), such as grammar induction (Meng and Siu, 2002) and affective text categorization (Malandrakis et al., 2011). Distributional semantic models (DSM) (Baroni and Lenci, 2010) are based on the distributional hypothesis of meaning (Harris, 1954) assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSM are typically constructed from co-occurrence statistics of word tuples that are extracted from a text corpus or from data harvested from the web. A wide range of contextual features are also used by DSM exploiting lexical, syntactic, semantic, and pragmatic information. DSM have been successfully applied to the problem of semantic similarity computation. According to (Baroni and Lenci, 2010), the success of contextual DSM features is due to their ability to encode the attributes of word senses. According to *attributional similarity* (Turney, 2006), semantic similarity between words is based on the commonality of their sense attributes. A closely related assumption is that the semantic similarity of two words can be estimated as the similarity of their two closest senses (Resnik, 1995), henceforth, referred to as the *maximum sense similarity* assumption.

In this paper, we investigate a new unsupervised approach for the construction of DSM with application to lexical semantic similarity computation. First, a corpus of snippets (short pieces of text containing words of interest) is harvested from the web. Then, a semantic network is constructed encoding the semantic relations between words in the corpus. Co-occurrence and context features are used to measure the strength of relations. The network is a parsimonious representation of the information encoded in the corpus. We then define the notion of semantic neighborhood and associated metrics of semantic similarity that exploit this notion. The proposed semantic similarity metrics are motivated by the maximum sense similarity, attributional similarity and metric space assumptions. The similarity metrics are evaluated against human similarity ratings using standard datasets, achieving state-of-the-art results. This work builds upon our prior research in (Iosif and Potamianos, 2010; Iosif and Potamianos, 2012b), while the following are the original contributions:

1. An efficient and scalable methodology is proposed, for corpus creation using web-harvested data. Unlike the quadratic query complexity of our previous algorithm (Iosif and Potamianos, 2010), the proposed method has linear query complexity with respect to the size of the lexicon.
2. Three unsupervised language-agnostic similarity computation algorithms are proposed that exploit the semantic neighborhoods. The best performing neighborhood-based metrics outperform well-established approaches that rely on elaborate knowledge resources.
3. We demonstrate the effectiveness of co-occurrence-based similarity metrics when corpus-based frequencies are incorporated in comparison to the use of web hits (Iosif and Potamianos, 2010). This is further investigated with respect to the textual proximity of co-occurring words.
4. The assumption that the semantic similarity between two words can be estimated as the similarity of their two closest senses is validated using (sense-untagged) web data.
5. The computation of semantic neighborhoods introduced in (Iosif and Potamianos, 2012b) is extended by applying a number of co-occurrence-based similarity metrics in addition to the context-based metrics. Word co-occurrence is shown to be more salient than contextual features regarding the discovery of senses via semantic neighborhoods.

The remainder of the work is organized as follows: In Section 2, we review related work in the areas of semantic similarity computation and word sense disambiguation. In Section 3 co-occurrence and context-based similarity metrics are reviewed. The procedure and motivation behind harvesting a corpus of snippets from the web is detailed in Section 4. In Section 5, we define our semantic network and propose three novel similarity metrics that utilize the notion of semantic neighborhood. The corpora and experimental procedures are described in Section 6, while the evaluation results are reported in Section 7. Last, Section 8 concludes this work.

2 Related Work

Semantic similarity metrics can be divided into two broad categories: (i) metrics that rely on knowledge resources, and (ii) corpus- or web-based metrics that do not require any external knowledge source. A representative example of the first category are metrics that exploit the WordNet ontology (Miller, 1990). For computing word similarity these metrics incorporate features such as the length of paths between them (Wu and Palmer, 1994; Leacock and Chodorow, 1998) or the information content of their least subsumer that is estimated from a corpus (Resnik, 1995; Jiang and Conrath, 1997). WordNet glosses have been also exploited for extracting contextual information (Banerjee and Pedersen, 2002; Patwardhan and Pedersen, 2006). An in depth review of the major WordNet-based metrics can be found in (Budanitsky and Hirst, 2006). Corpus-based metrics are formalized as DSM (Baroni and Lenci, 2010) and are based on the distributional hypothesis of meaning (Harris, 1954). DSM can be categorized into unstructured (unsupervised) that employ a bag-of-words model (Iosif and Potamianos, 2010) and structured that rely on syntactic relationships between words (Grefenstette, 1994; Baroni and Lenci, 2010).

The core idea behind structured models is the utilization of syntactic relationships as features for the creation of semantic spaces. Typical examples of such relations are argument structures (subject/object) and modifications (adjective-noun) extracted by shallow or full parsing (Padó and Lapata, 2007). Syntactic relations can be represented as 2-tuples of the arguments (Grefenstette, 1994) or as n-tuples in order to incorporate direct and indirect dependencies (Padó and Lapata, 2007). The paradigm of "one task, one model" of structured DSM was advanced in (Baroni and Lenci, 2010) by the arrangement of tuples into a third-order tensor. This enables the creation of different semantic spaces for different semantic tasks (e.g., estimation of semantic similarity between words, categorization of concepts, computation of verb selectional preferences, etc.), while the extraction of dependency tuples is task-independent ("the same distributional information can be shared across tasks"). The performance of knowledge-based metrics and DSM for the problem of semantic similarity estimation between words was compared in (Agirre et al., 2006). Unstructured DSM were shown to obtain slightly higher performance than structures ones, where knowledge-based and DSM approaches yielded comparable results. The best performance was obtained using a supervised combination of a knowledge-based (WordNet) approach and two variations of unstructured DSM.

Exemplar models were proposed as an alternative implementation of DSM for addressing the problem of polysemy (Erk and Padó, 2010; Reddy et al., 2011). Instead of a single vector, a set of exemplars is utilized for the representation of a target word. The set of exemplars is defined as the set of (corpus) sentences in which the target occurs (Erk and Padó, 2010). Each exemplar can be represented as a structured (i.e., bag-of-words) or unstructured (i.e., encoding syntactic relations) vector. For a target word given within context (e.g., sentential) polysemy is modeled by the activation/selection of the relevant exemplars with respect to a "point of comparison", where the latter can be regarded as another exemplar (Erk and Padó, 2010). A related example is the selection of paraphrases for a target word that occurs in a given context (Erk and Padó, 2010). In (Reddy et al., 2011), the exemplar model was applied in the framework of semantic compositionality for the task of similarity computation between noun-noun compounds.

Web-based metrics employ search engines to estimate the frequency of word co-occurrence (Vitanyi, 2005; Gracia et al., 2006; Turney, 2001) or construct corpora (Bollegala et al., 2007; Iosif and Potamianos, 2010). The identification and extraction of other types of relations has been mainly studied through the use of linguistic patterns. Lexico-syntactic patterns were applied in the influential work of Hearst (Hearst, 1992), for the identification of hyponymy, followed by numerous similar approaches, e.g., (Caraballo, 1999).

Recently, motivated by the graph theory, several aspects of the human languages have been modeled using network-based methods. In (Radev and Mihalcea, 2008; Mihalcea and Radev, 2011), an overview of network-based approaches is presented for a number of NLP problems. Different types of language units can be regarded as vertices of such networks, spanning from single words to sentences. Typically, network edges represent the relations of such units capturing phenomena such as co-occurrence, syntactic dependencies, and lexical similarity. An example of a large co-occurrence network is presented in (Widdows and Dorow, 2002) for the automatic creation of semantic classes. In (Ferrer-I-Cancho and Solé, 2001), it is reported that the co-occurrence networks of words that co-exist at very short proximity, exhibit a number of small-world properties and are highly clustered. Similar observations regarding the structural properties of co-occurrence networks were also made in (Véronis, 2004), where the HyprLex algorithm was proposed for sense discovery. In (Agirre et al., 2006), an extension of the main ideas presented in (Véronis, 2004) was proposed for word sense disambiguation (WSD). In particular, the PageRank algorithm (Brin and Page, 1998) was employed for identifying hubs over a co-occurrence network.

Semantic similarity computation is closely related to WSD. WSD methods can be divided into two main categories: (i) supervised approaches that apply machine learning for learning sense labels for a set of words with respect to a given context (sense labeling), and (ii) unsupervised approaches that automatically discriminate (discover) word senses without label assignment. A detailed survey of WSD is provided in (Ide and Véronis, 1998; Navigli, 2009; Agirre and Edmonds, 2007). The employment of network-based metrics for the computation of semantic similarity has attracted less attention compared to WSD. WordNet-based similarity metrics can be regarded as a special case of network metrics, since they are built on the top of a manually created network. To the best of our knowledge few network-based metrics are reported in the literature that integrate network creation with semantic similarity computation. In (Lemaire and Denhière, 2004), a co-occurrence network was constructed, and the similarity between two words was estimated as the product of weights of the shortest path between them with moderate performance results.

Other network-based approaches rely on the use of linguistic tools and/or knowledge resources. In (Harrington, 2010), a semantic network was created from a corpus using a set of tools for named entity recognition, parsing, and semantic analysis. The similarity between words (nodes) was estimated according to an implementation of the spreading activation model (Collins and Loftus, 1975). Wikipedia articles were exploited by the WikiRelate! system (Strube and Ponzetto, 2006) in conjunction with the corresponding category tree for the creation of a semantic network. In (Hughes and Ramage, 2007), an approach based on random walks was applied over WordNet for estimating word similarity. WordNet was also used in (Agirre et al., 2006) where the personalized PageRank algorithm (Haveliwala et al., 2002) was applied for the computation of a probability distribution for every target word. Word similarity was estimated via the cosine similarity between the vectorized distributions.

Following the paradigm of the vector space model (VSM) that constitutes the main implementation of DSMs, our approach is based on corpus-based co-occurrence statistics for the creation of a semantic network. One important difference with prior work in this area is that no language-specific tools, e.g., dependency parsers (Baroni and Lenci, 2010), or human annotations, e.g., Wikipedia hyperlinks (Wojtinnek et al., 2012), are used here. For example, in (Wojtinnek et al., 2012) the English Wikipedia was used for the disambiguation of target words (Wikipedia concepts) and a very large network was constructed by exploiting the hyperlinks between them. For each node (word) a vector was created including a number of strongly connected nodes selected by an algorithm inspired by spreading activation theory (Collins and Loftus, 1975). The similarity between two words was estimated as the cosine of their respective vectors. In our work, two types of metrics are investigated for weighting the strength of the link between a reference noun and its neighbors, namely, co-occurrence-based and context-based. Co-occurrence-based metrics were previously used for the weighting of contextual features (Agirre et al., 2009; Baroni and Lenci, 2010) and the creation of co-occurrence networks (Widdows and Dorow, 2002). To the best of our knowledge context-based metrics have never been applied for any of the aforementioned tasks. Our work is also motivated by cognitive consideration and theories of semantics. The network-based metrics proposed here are motivated by two well-founded hypotheses regarding semantic similarity, namely, maximum sense similarity (Resnik, 1995) and attributional similarity (Turney, 2006). Our work extends the traditional VSM approach into a two tier system: corpus statistics are parsimoniously encoded in a network, while the task of similarity computation is shifted (from corpus-based techniques) to operations over network neighborhoods. The proposed network creation process constitutes a new paradigm for implementing DSMs that enables the direct exploitation of neighborhood semantics, e.g., definition of metrics that adopt different hypotheses regarding semantic similarity, investigation of neighborhood structural properties.

3 Similarity Metrics

Next, the two main types of similarity metrics used in this paper are presented, namely, co-occurrence and context-based metrics.

3.1 Co-occurrence-based Metrics

The underlying assumption of co-occurrence-based metrics is that two words that co-exist in the same web document are semantically related. Let $\{M\}$ be the entire set of web documents indexed by a search engine. A set of documents returned by a search engine for query words $w_i, \ldots, w_{i+n}$ is denoted as $\{M; w_i, \ldots, w_{i+n}\}$, with cardinality $|\,M; w_i, \ldots, w_{i+n}\,|$. The latter is also known as the number of hits for the submitted query. In this work, we employ four co-occurrence-based similarity metrics defined next (Iosif and Potamianos, 2010).

Jaccard coefficient: In general, this coefficient computes the similarity between sets. In our case, we consider the sets of web documents that are indexed by the words of interest. The Jaccard coefficient J between words w_i and w_j is defined as follows:

$$J(w_i, w_j) = \frac{|\,M; w_i, w_j\,|}{|\,M; w_i\,| + |\,M; w_j\,| - |\,M; w_i, w_j\,|}. \tag{1}$$

The Jaccard coefficient takes value between 0 (totally dissimilar) and 1 (totally similar).

Dice coefficient: This coefficient is closely related to the Jaccard coefficient and it is defined as:

$$D(w_i, w_j) = \frac{2 \mid M; w_i, w_j \mid}{\mid M; w_i \mid + \mid M; w_j \mid}. \tag{2}$$

As before, D ranges between 0 and 1 for absolute dissimilarity and similarity, respectively.

Mutual information: Assuming that $\mid M; w_i \mid$, $\mid M; w_j \mid$ are random variables, then their pointwise mutual information reflects the dependence between the occurrence of w_i and w_j, as follows (Bollegala et al., 2007):

$$I(w_i, w_j) = \log \frac{\frac{\mid M; w_i, w_j \mid}{\mid M \mid}}{\frac{\mid M; w_i \mid}{\mid M \mid} \frac{\mid M; w_j \mid}{\mid M \mid}}. \tag{3}$$

The mutual information is unbounded ranging from positive to negative values as the similarity of words decreases.

Google-based semantic relatedness: The “normalized Google distance” was proposed in (Vitanyi, 2005), defined as follows:

$$G_0(w_i, w_j) = \frac{\max\{\log \mid M; w_i \mid, \log \mid M; w_j \mid\} - \log \mid M; w_i, w_j \mid}{\log \mid M \mid - \min\{\log \mid M; w_i \mid, \log \mid M; w_j \mid\}}. \tag{4}$$

This metric is a dissimilarity measure, i.e., as the semantic similarity between two words increases the metric takes smaller values. The scores assigned by (4) are unbounded, ranging from 0 to ∞. In (Gracia et al., 2006), a variation of the normalized Google distance was used, proposing a bounded similarity measure called “Google-based semantic relatedness”, defined as:

$$G(w_i, w_j) = e^{-2G_0(w_i, w_j)}, \tag{5}$$

$G_0(w_i, w_j)$ is computed according to (4). The Google-based semantic relatedness is bounded in [0,1].

Beside the use of the web as a corpus for obtaining number of hits, the aforementioned co-occurrence-based metrics can be also defined with respect to any text corpus. In such cases, the word frequencies can be considered at the level of several corpus units, e.g., sentences, paragraphs. In this work, we adopt both perspectives, i.e., employing web-based hits, as well as word frequencies computed over a text corpus (see Section 7.1). Beyond computing similarities at the word level, the proposed metrics have been also shown to perform well for phrase-level similarity tasks. In (Malandrakis et al., 2012), a parametric version of mutual information using the corpus creation method proposed here was integrated in a sentence-level model with good results.

3.2 Context-based Metrics

The fundamental assumption behind context-based metrics is that *similarity of context implies similarity of meaning*: we expect that words that share similar lexical contexts will be semantically related, (Harris, 1954). A common representation of contextual features is the "bag-of-words" model that assumes independence between features (Sebastiani and Ricerche, 2002).

For context-based metrics, a contextual window of size $2H + 1$ words is centered on the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus, the H words left and right of w_i are taken into consideration, i.e.,

$$[f_{H,l} \;...\; f_{2,l} \;\; f_{1,l}] \;\; w_i \;\; [f_{1,r} \;\; f_{2,r} \;...\; f_{H,r}],$$

where $f_{k,l}$ and $f_{k,r}$ represent the feature k positions to the left of right of w_i. For a given value of H, the feature vector for w_i is built as $T_{w_i,H} = (t_{w_i,1}, t_{w_i,2}, ..., t_{w_i,Z})$, where $t_{w_i,k}$ is a non-negative integer. The feature vector has length equal to the vocabulary size Z. Non-zero feature value $t_{w_i,k}$ indicates the occurrence of vocabulary word t_k within the left or right context of w_i. Note that the value of $t_{w_i,k}$ is set by considering all occurrences of w_i in the corpus. The value of $t_{w_i,k}$ can be defined according to a binary scheme [1] (Iosif and Potamianos, 2010). This scheme assigns 1 to $t_{w_i,k}$ if vocabulary word t_k occurs within H positions left or right of word w_i, otherwise, $t_{w_i,k} = 0$. The context-based semantic similarity metric Q^H between words w_i and w_j is computed as the cosine of their feature vectors:

$$Q^H(w_i, w_j) = \frac{\sum_{k=1}^{Z} t_{w_i,k} \; t_{w_j,k}}{\sqrt{\sum_{k=1}^{Z} t_{w_i,k}^2} \sqrt{\sum_{k=1}^{Z} t_{w_j,k}^2}}, \tag{6}$$

for context size H and vocabulary size Z. For words w_i, w_j that share no common context (completely dissimilar words) the corresponding semantic similarity score is 0. Also $Q^H(w, w) = 1$.

4 Corpus Creation Using Targeted Web Queries

In this section, we investigate the creation of corpora from web-harvested data via the formulation of targeted queries. There are two main types of web queries that can be used for corpus creation: (i) conjunctive queries (AND), and (ii) individual queries (IND) [2] . Assuming N words in our lexicon, in the first case all pairwise AND conjunctions are formed and the corresponding queries are posed to a web engine, e.g., "w_i AND w_j". Corpus creation via AND queries leads to quadratic query complexity $\mathcal{O}(N^2)$ in the number of words in the lexicon. Alternatively, one can download documents or snippets with linear query complexity $\mathcal{O}(N)$ using IND queries, i.e., "w_i".

The main advantage of AND queries is that they construct a corpus that is conditioned on word-pairs, explicitly requesting the co-occurrence of word-pairs in the same document.

[1] In (Iosif and Potamianos, 2010), the binary scheme perfromed the best among the various contextual weighting schemes investigated for a word-level semantic similarity task. For this task, similar performance can be achieved using the χ^2 statistic (Agirre et al., 2009). A detailed evaluation of the various context weighting schemes is beyond the scope of this paper, since it does not affect the main conclusions drawn on semantic network construction and metric evaluation.

[2] Word co-occurrence statistics estimated on a web-harvested corpus may be biased due to optimizations applied by web search engines when ranking documents and selecting snippets. This is especially true for query words in corpora resulting from conjunctive AND queries, but less so for corpora harvested via IND queries.

Co-occurrence is a strong indicator of similarity and corpora created via AND queries have been shown to provide very good semantic similarity estimates (Iosif and Potamianos, 2010). To better understand the role of co-occurrence as a feature in semantic similarity computation, we need to revisit the very definition of semantic similarity, as it pertains to words and their senses. According to the information-theoretic approach proposed in (Resnik, 1995), the similarity of two concepts can be estimated as the similarity of their two closest senses. This is also in agreement with our "common sense" (cognitive) model of semantic similarity: when two words are mentioned, their closest senses are activated[3]. We believe that an important contribution of the co-occurrence feature to semantic similarity computation is that *co-occurrence acts as a semantic filter that only retains the two closest senses*. See Section 7.2 for the experimental justification of this claim.

Unfortunately attempting to build corpora and DSM using conjunctive AND queries does not scale to thousands of words due to quadratic query complexity [4]. We are thus forced to investigate the alternative of using IND queries and face the sense disambiguation issues associated with such corpora. Corpora created via IND queries are similar to a typical text corpus with one important difference: the frequency of occurrence of the words in our lexicon can be manipulated to deviate from Zipf's law. Assuming that the same number of snippets is downloaded for each word in our lexicon (using IND queries), we expect that rare words will be well-represented within the corpus. As a result, the corpus will be more "informative", i.e., the entropy rate of a unigram (zeroth order Markov process) model will be higher.

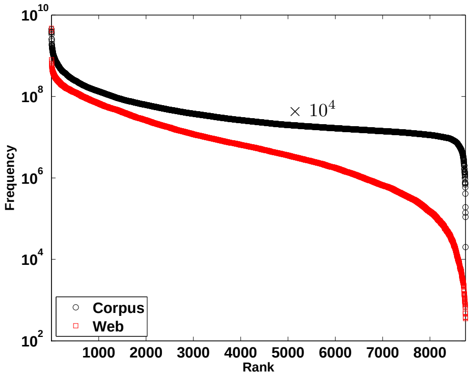


Fig. 1. Frequency of $8,752$ nouns vs. their rank. The frequencies were computed using 1) corpus counts (black curve), and 2) web hits (red curve). For comparison purposes the corpus frequencies were multiplied by 10^4.

The normalization word-frequency effect can be illustrated by plotting the empirical distribution of the frequency of the words in the lexicon. Using a lexicon of $8,752$ nouns, the noun frequencies are plotted as a function of their rank in Fig.1. More specifically, we

[3] The maximum sense similarity assertion is widely employed by many top-performing similarity metrics, such as the WordNet-based metrics (Budanitsky and Hirst, 2006).

[4] Although a work-around could be found, e.g., using cross-products of all term statistics in a search engine index and n-gram counts.

created a web corpus by posing an IND query for each noun and retrieving the $1,000$ top-ranked snippets (see Section 6). The corpus frequencies were multiplied by 10^4 in order to facilitate the comparison with the red curve showing web hits frequency. According to the Zipf's law (Zipf, 1965), the frequency of a word w decreases non-linearly as its rank increases:

$$f(w) = \frac{c}{r(w)^{\gamma}}, \tag{7}$$

where $f(w)$ and $r(w)$ are the frequency and the rank of word w, respectively, while c and γ are corpus-dependent. It is clear that the frequency difference between the high-ranked and the low-ranked words is somewhat normalized, i.e., smaller (absolute) γ, for the case of corpus frequencies, as opposed to the use of web hits. For the example of Fig.1, γ equals to -0.54 and -0.90 for corpus frequencies and web hits, respectively. These values were computed for the ranks lying between $1,000$ and $6,000$ using a least squares linear model. This normalization is expected to smooth the domination of very frequent words at the denominator of co-occurrence-based metrics, such as (1)–(3). The performance of web hits and corpus counts is presented in Table 2.

5 Semantic Network

Next, we construct a semantic network encoding the relevant corpus statistics. The network is defined as an undirected (under a symmetric similarity metric) graph $F = (V, E)$ whose the set of vertices V are all words in our lexicon L, and the set of edges E contains the links between the vertices. The links (edges) between words in the network are determined and weighted according to the pairwise semantic similarity of the vertices.

The network is a parsimonious representation of corpus statistics as they pertain to the estimation of semantic similarities between word-pairs in the lexicon. In addition, the network can be used to *discover relations that are not directly observable in the data*; such relations emerge via the systematic covariation of similarity metrics. Semantic neighborhoods play an important role in this process. The members of the semantic neighborhoods of words are expected to contain features capturing diverse information at the syntactic, semantic and pragmatic level.

5.1 Semantic Neighborhoods

For each word (reference word) that is included in the lexicon, $w_i \in L$, we consider a subgraph of F, $F_i = (N_i, E_i)$, where the set of vertices N_i includes in total n members of L, which are linked with w_i via edges E_i. The F_i subgraph is referred to as the semantic neighborhood of w_i (Iosif and Potamianos, 2012b). The members of N_i (neighbors of w_i) are selected according to a semantic similarity metric (co-occurrence-based defined in Section 3.1, or context-based defined in Section 3.2) with respect to w_i, i.e., the n most similar words to w_i are selected. Note that the semantic network is not a metric space under (the proposed co-occurrence or context-based) semantic similarity because the triangle inequality is, in general, not satisfied. Next, we propose three semantic similarity metrics that utilize the notion of semantic neighborhood.

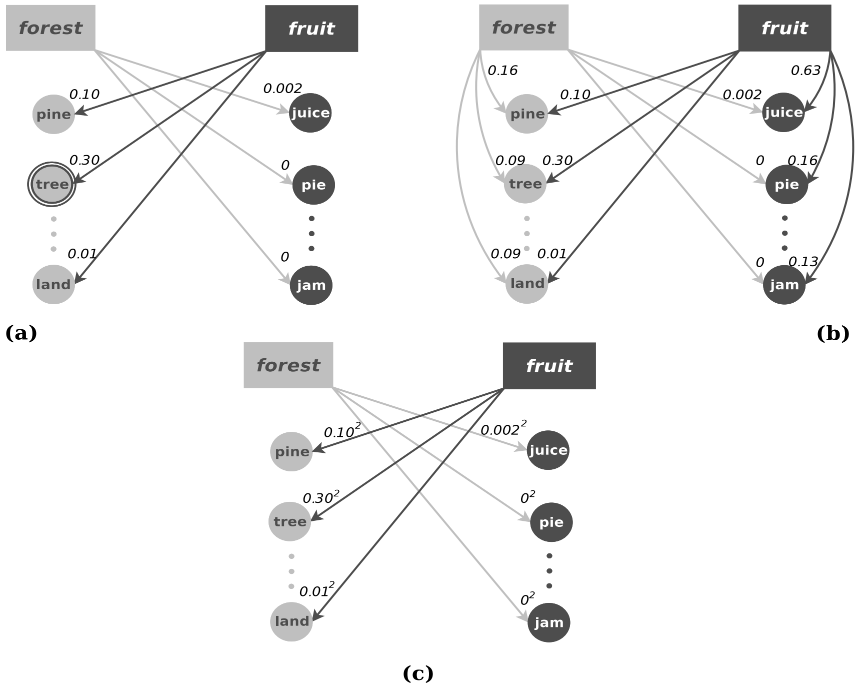


Fig. 2. Pictorial view of neighborhood-based metrics. Two reference nouns, "forest" and "fruit", are depicted along with their neighborhoods: {pine, tree, . . . , land} and {juice, pie, . . . , jam}, respectively. Arcs represent the similarities between reference nouns and neighbors. The similarity between "forest" and "fruit" is computed according to (a) maximum similarity of neighborhoods, (b) correlation of neighborhood similarities, and (c) sum of squared neighborhood similarities.

5.2 Maximum Similarity of Neighborhoods

This metric is based on the hypothesis that the similarity of two words, w_i and w_j, can be estimated by *the maximum similarity of their respective sets of neighbors*, defined as follows:

$$M_n(w_i, w_j) = \max\{\alpha_{ij}, \alpha_{ji}\}, \tag{8}$$

where

$$\alpha_{ij} = \max_{x \,\in\, N_j} S(w_i, x), \quad \alpha_{ji} = \max_{y \,\in\, N_i} S(w_j, y).$$

α_{ij} (or α_{ji}) denotes the maximum similarity between w_i (or w_j) and the neighbors of w_j (or w_i) that is computed according to a similarity metric S (see for example (1)–(3), (5), (6)). N_i and N_j are the set of neighbors for w_i and w_j, respectively. The definition of M_n is motivated by the maximum sense similarity assumption[5]. As discussed above, semantic neighborhoods encode diverse information. Here the underlying assumption is that the most salient information in the neighbors of a word are semantic features denoting senses

[5] This metric utilizes the similarities between w_i and x, $\forall\, x \in N_j$, as well as between w_j and y, $\forall\, y \in N_i$. This is slightly different than considering all the pairwise similarities between the members of N_i and N_j.

of this word. In other words, we assume that semantic neighborhoods (and semantic networks, in general) can be used to mine for word senses[6]. The M_n metric takes values in the interval $[0, 1]$, where 1 stands for absolute similarity. Also, $M_n(w_i, w_j) = M_n(w_j, w_i)$, i.e., M_n is symmetric. An example illustrating the computation of similarity between "forest" and "fruit" is depicted by Fig.2(a). $M_n($"*forest*", "*fruit*"$) = 0.30$ because the similarity between "fruit" and "tree" (among all neighbors of "forest") is the largest.

5.3 Correlation of Neighborhood Similarities

The similarity between w_i and w_j is defined as follows:

$$R_n(w_i, w_j) = \max\{\beta_{ij}, \beta_{ji}\}, \tag{9}$$

where

$$\beta_{ij} = \rho(C_i^{N_i}, C_j^{N_i}), \ \ \beta_{ji} = \rho(C_i^{N_j}, C_j^{N_j})$$

and

$$C_i^{N_i} = (S(w_i, x_1), S(w_i, x_2), \ldots, S(w_i, x_n)), \quad \text{where} \ \ N_i = \{x_1, x_2, \ldots, x_n\}.$$

Note that $C_j^{N_i}$, $C_i^{N_j}$, and $C_j^{N_j}$ are defined similarly as $C_i^{N_i}$. The ρ function stands for the Pearson's correlation coefficient, N_i is the set of neighbors of word w_i, and S is a similarity metric. Here, we aim to exploit the entire semantic neighborhoods for the computation of semantic similarity, as opposed to M_n where a single neighbor is utilized. The motivation behind this metric is attributional similarity, i.e., we assume that semantic neighborhoods encode attributes (or features) of a word. Neighborhood correlation similarity in essence compares the distribution of semantic similarities of the two words on their semantic neighborhoods. Thus, this metric is expected to provide more robust similarity estimates compared to M_n, especially when few data are available. The ρ function incorporates the covariation of the similarities of w_i and w_j with respect to the members of their semantic neighborhoods. The underlying assumption is that two semantically similar words are expected to have co-varying similarities with respect to their neighbors. Moreover, the ρ function normalizes this covariance by the standard deviations of the similarities of w_i and w_j. The similarity scores computed by R_n metric ranges in the interval $[-1, 1]$, where -1 and 1 denote zero and absolute similarity, respectively. R_n is symmetric, since $R_n(w_i, w_j) = R_n(w_j, w_i)$. The similarity computation process is exemplified in Fig.2(b) for the words $w_1 =$"forest" and $w_2 =$ "fruit". The similarity vectors between the neighbors N_1 of "forest" and each of the words are computed: $C_1^{N_1} = (0.16, 0.09, \ldots, 0.09)$, $C_2^{N_1} = (0.10, 0.30, \ldots, 0.01)$. Similarly, $C_1^{N_2}$, $C_2^{N_2}$ are computed for the neighbors of "fruit" and combined to estimate $R_n($"*forest*", "*fruit*"$) = -0.04$.

[6] See also (Navigli and Crisafulli, 2010) for word sense discovery via semantic networks.

5.4 Sum of Squared Neighborhood Similarities

The similarity between w_i and w_j is defined as follows:

$$E_n^\theta(w_i,w_j)=\left(\sum_{x\,\in\,N_j} S^\theta(w_i,x)+\sum_{y\,\in\,N_i} S^\theta(w_j,y)\right)^{\frac{1}{\theta}}, \tag{10}$$

where N_i is the set of neighbors of word w_i, and S is any similarity metric. Similar to (9) all neighbors contribute to the computation of the final similarity score, here this is performed by summing the squares ($\theta = 2$) of similarities between w_i and w_j's neighbors. The same calculation is repeated for w_j and the neighbors of w_i to make $E_n^\theta(w_i,w_j)$ symmetric. This is illustrated by Fig.2(c) for the computation of similarity between "forest" and "fruit" for $\theta = 2$. That is $E_n^{\theta=2}(\textit{“forest”}, \textit{“fruit”}) = \sqrt{(0.10^2+0.30^2+\cdots+0.01^2)+(0.002^2+0^2+\cdots+0^2)}$ $= 0.22$. The $E_n^{\theta=2}$ metric is unbounded since the yielding similarity scores range within $[0,\infty)$. This range is smoothed in a non-linear way by taking the square root of the accumulated squares of similarities. As in (9), the motivation underlying $E_n^{\theta=2}$ metric is the attributional similarity, i.e., neighbors stand as attributes (or features). However, what is different here is the utilization of the attributional similarity as indicator for semantic similarity, i.e., the accumulation of word–to–neighbor similarities. The contribution of each word–to–neighbor similarity is non-linearly weighted using the square of the respective similarity score. The motivation behind using $\theta > 1$ is that more similar words in the neighborhoods should be weighted more in the final similarity decision[7]. Qualitatively, the $E_n^{\theta=2}$ weighting scheme takes the middle road between selecting the maximum pairwise similarity in (8) and the "linear" weighting of pairwise similarity in (9). Note that as θ goes to ∞, E_n^θ and M_n become equivalent.

6 Evaluation Datasets, Corpora and Experimental Procedure

6.1 Evaluation Datasets

The performance of similarity metrics was evaluated against human ratings from three standard datasets of noun pairs, namely: 1) MC (Miller and Charles, 1998), 2) RG (Rubenstein and Goodenough, 1965), and 3) WS353 (Finkelstein et al., 2002). The first dataset consists of 28 noun pairs. For the second and the third dataset we present results for the subset of 57 and 272 pairs, respectively, that are also included in SemCor3[8] corpus. The Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against the ground truth. Let $X = (x_1, x_2, ..., x_m)$ and $Y = (y_1, y_2, ..., y_m)$ be the vectors that contain the similarity scores given by human subjects and the computational metric, respectively, for each of the $i = 1, 2, ..., m$ word pairs of the datasets. Pearson's

[7] Despite the resemblance between the $E_n^{\theta=2}$ metric and the Euclidean distance, no assumption is adopted here about the semantic neighborhoods being metric spaces under S.

[8] `http://www.cse.unt.edu/~rada/downloads.html`

correlation coefficient is computed as follows:

$$\rho_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, for $i = 1, 2, ..., m$. This coefficient was selected instead of Spearman's rank correlation coefficient in order to retain the initial scaling of similarities in the evaluation metric, as opposed to the alternation of this scaling through the transformation of similarities into ranks.

6.2 Experimental Corpora and Procedure

We created the following corpora of web snippets using AND or IND queries posed via the Yahoo! Search API. 1) Corpus1: Using AND queries $1,000$ snippets were acquired for each pair of nouns, for the MC dataset. The major aspect of this corpus is the (explicitly requested) co-occurrence of nouns for which the similarity is computed. 2) Corpus2: Using IND queries $1,000$ snippets were acquired for each (unique) noun of the MC dataset. Unlike Corpus1, the creation of Corpus 2 is not driven by the co-occurrence constraint. 3) Corpus3: The same IND queries were used as for the case of Corpus2, but the queries were augmented with lexical descriptors denoting senses (see Section 7.2 for details). Corpus3 can be regarded as an extension of Corpus2, in which the acquired data are intended to (uniformly) cover the different senses of nouns. The aforementioned corpora are exploited (see Section 7.2) for investigating the effect of word co-occurrence and their senses to the computation of context-based similarity (for the MC dataset only). 4) Corpus4: This is a corpus created using IND queries, consisting of approximately $8,752,000$ snippets. More specifically, $1,000$ snippets were acquired for each noun taken from a set of $8,752$ English nouns of the SemCor3 corpus. Corpus4 is used for the creation of the semantic network as described in Section 5.

For Corpus4 the baseline performance of co-occurrence and context-based similarity metrics was computed (see also below for parameter definition). Then the semantic neighborhoods were defined and the maximum/correlation neighborhood similarities were computed. A detailed list of experiments was conducted trying to investigate the performance of the following list of parameters: 1) the size of the contextual window, H, used in Q^H, M_n, R_n, $E_n^{\theta=2}$ 2) the metric used for the selection of neighbors: co-occurrence-based (J, D, I, G) or context-based similarity (Q^H), 3) the S metric used in M_n, R_n, $E_n^{\theta=2}$: co-occurrence-based (J, D, I, G) or context-based similarity (Q^H), 4) the neighborhood size (number of neighbors n), used in M_n, R_n, $E_n^{\theta=2}$ metrics, 5) the corpus size, i.e., number of snippets per word (50, 100, 200, 500, 1,000) used to construct the network, and 6) the network size, that is the number of concepts (nouns of lexicon) that constitute the network: $9,\ 88,\ 176,\ 876,\ 1,751,\ 4,376,\ 6,127,$ and $8,752$. The results are presented next.

7 Results

The performance of the context-based metric and the co-occurrence-based metrics defined in Section 3 is compared in Section 7.1 (baseline performance). In Section 7.2, we compare the performance of the baseline context-based metric for corpora created via AND and IND queries, and we show that senses play an important role in achieving good performance. In Section 7.3, we present the performance of the proposed neighborhood-based metrics, defined in Section 5, that utilize the large corpus created via IND queries (Corpus4) and the corresponding semantic network.

7.1 Baseline

We consider as baseline the performance of the following metrics: 1) context-based similarity metric Q^H defined in Section 3.2, 2) co-occurrence-based metrics, defined in Section 3.1, relying on counts that were computed either using the web as a corpus (number of hits), or the corpus of snippets harvested with respect to the $8,752$ nouns (Corpus4). The baseline scores for the context-based similarity metric Q^H are presented in Table 1, for several values of the contextual window size H. The best correlation scores are obtained for $H=1$ across all datasets, while the performance drops as the size of the contextual window increases. Even for $H=1$, moderate correlation scores are achieved for the MC and RG datasets, while the baseline performance is poor for the WS353 dataset. These results indicate the inability of naive context-based similarity metrics to exploit contextual features, despite the availability of a large corpus. The baseline performance for co-occurrence-

Table 1. *Performance of context-based metric Q^H for several values of H.*

Dataset	Contextual window size			
	$H=1$	$H=2$	$H=3$	$H=5$
MC	**0.53**	0.35	0.29	0.20
RG	**0.52**	0.41	0.37	0.29
WS353	**0.30**	0.21	0.17	0.13

Table 2. *Performance of co-occurrence-based metrics using web and corpus counts: Jaccard (J), Dice (D), Mutual info. (I), and Google-based sem. rel. (G).*

Dataset	Co-occurrence-based metrics using							
	Web counts				Corpus counts			
	J	D	I	G	J	D	I	G
MC	-0.20	0.24	**0.35**	0.33	0.59	0.59	0.78	**0.85**
RG	-0.01	0.21	**0.28**	0.31	0.60	0.60	0.77	**0.81**
WS353	-0.02	0.10	0.19	**0.20**	0.18	0.22	0.60	**0.61**

based metrics that incorporate web counts[9] (hits) or corpus counts (Corpus4) is shown in Table 2. Regarding corpus counts, the co-occurrence of nouns is considered at the snippet boundary. We observe that the employment of corpus counts leads to significantly higher correlation scores, compared to using web counts. For example, the correlation improves from 0.33 to 0.85 using the G metric for the case of the MC dataset. This observation is consistent for all metrics across all three datasets. For corpus counts, the best performance is achieved by Google-based Semantic Relatedness, G, while the Mutual information, I, is a close second. Jaccard, J, and Dice, D, coefficients have lower but comparable performance.

Table 3. *Percentage of highly related pairs that have zero co-occurrence corpus counts as a function of downloaded snippets.*

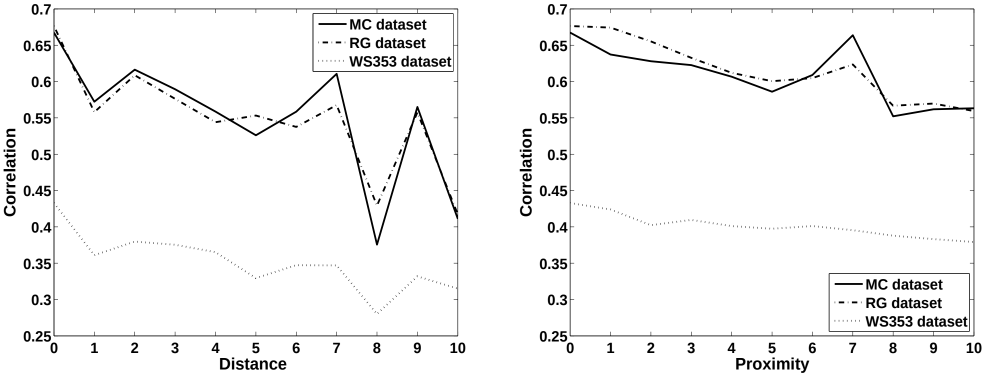
Dataset	Number of snippets				
	50	100	200	500	1000
MC	43%	28%	10%	3%	0%
RG	40%	24%	12%	8%	5%
WS353	20%	13%	8%	3%	3%

Despite the high performance of co-occurrence metrics using corpus counts, their applicability is strongly depended on the corpus size. The percentage of highly related noun pairs that have zero co-occurrence (corpus) counts are presented in Table 3, for several numbers of downloaded snippets. We assumed that two nouns are highly related if the corresponding similarity score (normalized between 0 and 1) provided by human subjects is greater than 0.5. The reported number of snippets were randomly selected from the initial (full) corpus. We observe that for the RG and WS353 datasets, even for the maximum number of snippets ($1,000$ per noun) 3–5% of the highly related pairs do not co-occur.

The poor performance of co-occurrence-based metrics that rely on web counts may be attributed to the fact that the co-occurrence of words is estimated at the document level, rather than at the level of snippet or sentence (for corpus counts). The key difference between web and corpus counts is the proximity of the co-occurring words, as well as, the different corpus statistics shown in Fig.1. In order to investigate the role of proximity we formulated NEAR queries that constrain the distance between two words in an AND web query[10]. The performance of the I metric using web counts is presented as a function of the distance and proximity of co-occurring words (within documents) in Fig.3(a) and Fig.3(b), respectively. The distance, δ, between two co-occurring words denotes that exactly δ tokens interfere between them. The proximity, π_δ, of two words allows to π tokens to appear

[9] The Exalead web search engine was used (`http://www.exalead.com/search/`).

[10] This was performed by using the NEAR operator which is supported by the Exalead search engine. For example, the "w_i NEAR/2 w_j" query returns the number of hits for which words w_i and w_j co-occur at proximity equal to 2.



(a) (b)

Fig. 3. Correlation performance of the co-occurrence-based metric I vs. word (a) distance and (b) proximity (within web documents).

between them, where $0 \leq \pi_\delta \leq \delta$. We observe that imposing a distance/proximity constraint significantly improves the achieved correlation compared to the baseline of web co-occurrence counts in Table 2. For example, the correlation for the RG dataset improves from 0.28 (see Table 2) to 0.68 for $\delta = 0$ and π_0. Despite the clear improvement in the performance of web-based count performance (when applying a proximity constraint), corpus-based counts still outperform web-based counts. The second reason behind the superior performance of corpus-based counts is the (normalized) word frequency statistics[11] of the snippet corpus (vs. web) shown in Fig. 1.

7.2 Incorporating Word Senses Through Web Queries

We compare the performance of context-based similarity metrics for web corpora created via AND or IND queries. All results reported in this section are for the MC dataset. Baseline similarity scores here are computed using the $Q^{H=1}$ metric[12] defined in (6). The correlation scores for the context-based similarity metric using AND and IND queries (dotted and solid line, respectively) are shown in Fig. 4 as a function of the number of snippets. The performance for AND queries is a single point and was obtained at $1,000$ queries[13] (shown here as reference). It is clear that context-based similarity metrics perform much better when using AND rather than IND queries.

Our hypothesis is that the very good performance of AND queries is due to co-occurrence acting as a semantic filter that retains the two closest senses of the two words. Moreover, the poor performance of IND queries is due to the limited coverage of senses

[11] For a theoretical analysis of how word-frequency normalization in a web snippet corpus reduces the estimation error of co-occurrence similarity metrics see (Iosif and Potamianos, 2012a).

[12] For the rest experiments in this paper we use a context window of $H = 1$. Our experiments, as well as, our prior work (Iosif and Potamianos, 2010) indicate that $H = 1$ provides the best results for the problem of similarity computation.

[13] The maximum number of IND queries is greater than the number of AND queries, due to the use of two individual queries, instead of a single conjunctive query. Web search engines return up to $1,000$ snippets per query.

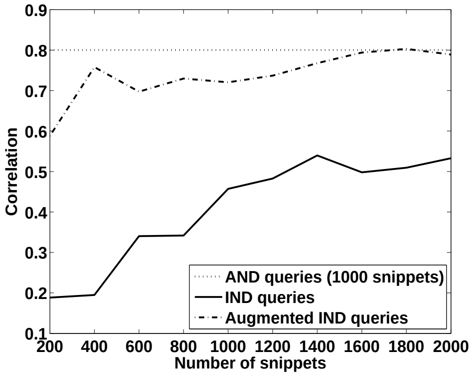


Fig. 4. Correlation performance for context-based similarity for web corpora created via AND queries (dotted), IND queries (solid), and IND queries augmented with sense descriptors (dashed–dotted) for the MC dataset.

within the top snippets. In order to verify this hypothesis we perform (sense) filtering explicitly following three steps: 1) identify all senses of the words of interest using WordNet[14], 2) use conjunctive AND queries between a word and each of its word senses to obtain relevant snippets that (mainly) contain the desired sense, e.g., the IND query for “magician” becomes “magician AND illusionist” (augmented), given the first WordNet sense of this word, and 3) compute the context-based similarity between all possible pairs of word senses and select the maximum similarity. Step 3 makes the implicit assumption that word similarity should be computed between the two closest senses (Budanitsky and Hirst, 2006), i.e., if s_{ik} is the kth sense of the word w_i the maximum sense context-based similarity Q' between words w_i, w_j is defined as:

$$Q'(w_i, w_j) = \max_{k,l} Q(s_{ik}, s_{jl}), \tag{11}$$

where Q is defined in (6). The performance of the augmented IND queries is shown in Fig.4 with a dashed-dotted line. It is clear that the use of the augmented IND queries significantly outperforms simple IND queries and approaches the performance of AND queries as the number of snippets increases.

Overall, the presented results suggest that the exploitation of word senses is essential for the accurate computation of semantic similarity. We have also experimentally demonstrated that context-based semantic similarity estimates are more accurate if we consider the two closest senses, i.e., the maximum pair-wise sense similarity score. A major roadblock in top-down corpus creation using IND queries is the lack of sense coverage in the corpus. We show next that by creating a corpus by posing IND queries for thousands of words, as well as, by employing the notion of semantic neighborhood we can overcome this roadblock and obtain excellent semantic similarity estimates.

[14] WordNet is used here simply to validate this hypothesis.

7.3 Semantic Network

Next, we investigate the computation of semantic networks using different types of similarity metrics. Next, we present the evaluation results for the proposed neighborhood-based similarity metrics, defined by (8)–(10), for different ways of defining the semantic neighborhoods.

7.3.1 Semantic Neighborhoods

The semantic neighborhood of each word is estimated using one of the co-occurrence-based metrics defined in Section 3.1, or the context-based similarity metric Q^H defined in Section 3.2. Our semantic network consists of $8,752$ nouns. Given a (reference) noun w, let $A(w)$ and $B(w)$ be the neighborhood sets of w computed using co-occurrence-based and context-based metrics. The intersection of $A(w)$ and $B(w)$, $A(w) \cap B(w)$, as well as their differences, $A(w) - B(w)$ and $B(w) - A(w)$, are shown in Table 4 for ten nouns that

Table 4. *Excerpts of semantic neighborhoods for ten nouns using the co-occurrence-based metric Dice (D) and/or the context-based metric $Q^{H=1}$.*

Reference	Neighbors selected by		
noun (w)	D and $Q^{H=1}$ $(A(w) \cap B(w))$	D only $(A(w) - B(w))$	$Q^{H=1}$ only $(B(w) - A(w))$
automobile	**auto**, **vehicle**, **car**, engine	accident, mechanic, starter, convertible	bus, aviation, tractor, lighting
brother	son, father, nephew, dad	twin, priest, police, girl	guy, lawyer, neighbor, pianist
car	**vehicle**, travel, service, price	accident, driver, **automobile**, fuel	business, city, game, quality
coast	island, **beach**, resort, sea	bay, boat, tsunami, port	lake, summer, entertainment, weather
food	water, health, service, industry	meal, kitchen, snack, gourmet	product, market, quality, life
forest	**land**, tree, vegetation, wildlife	rain, fire, pine, wood	nature, region, environment, property
fruit	tree, **plant**, taste, juice	vine, jam, acidity, pie	meal, wood, food, garden
hill	mountain, tree, park, forest	slope, **mound**, walk, snowball	island, city, resort, summer
journey	**trip**, destination, adventure, **travel**	discovery, quest, voyage, road	vision, goal, holiday, culture
slave	nigger, slavery, servant, manumission	gladiator, labor, freedom, master	beggar, democracy, society, aristocracy

are included in the experimental datasets. The co-occurrence-based metric D defined in (2) was applied for the computation of $A(w)$, while the context-based metric $Q^{H=1}$ defined in (6) was used for the computation of $B(w)$. For both metrics, the 50 top-ranked neighbors were considered. The neighbors that are emphasized using bold fonts denote (lexicalized) senses of the respective reference nouns.

We observe that the discovery of a number of senses via the neighborhoods is feasible for some nouns, e.g., “automobile” and “car”. This is more clear for $A(w) \cap B(w)$ compared to $A(w) - B(w)$ and $B(w) - A(w)$. However, sense discovery appears to be difficult for other nouns, such as “food” and “slave”, for which their respective senses can not be easily described by single words. In addition to synonymy, taxonomic relations are encoded by the neighbors of $A(w) \cap B(w)$, e.g., IsA(vehicle, car), PartOf(automobile, engine). Relations of associative nature, e.g., ProducedBy(industry, food), are also denoted by some neighbors of $A(w) \cap B(w)$. Essentially, the main difference between $A(w) - B(w)$ and $B(w) - A(w)$ is that the former includes members that tend to formulate more direct associative relations with the reference nouns. In some cases these relations appear in the corpus as bigrams, such as “car accident” and “hill slope”. Members of $B(w) - A(w)$ seem to correspond to relations of a broader semantic/pragmatic scope, such as (food, life) and (journey, culture).

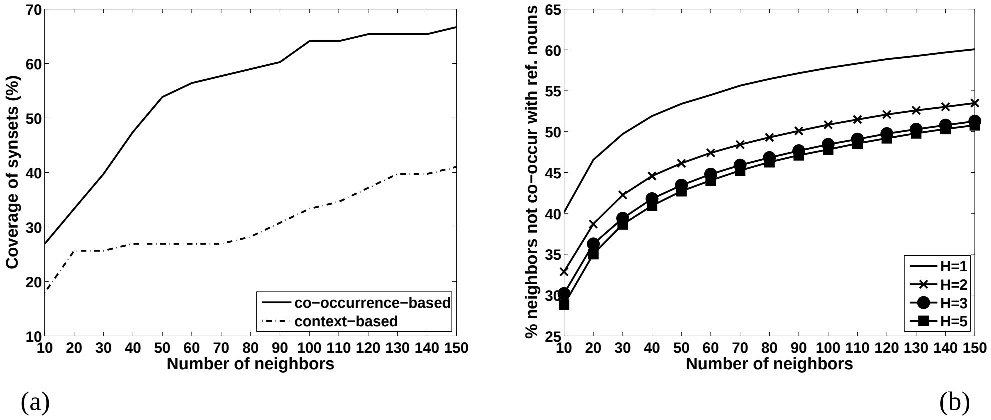


Fig. 5. (a) Percentage of WordNet synonyms included in the semantic neighborhoods vs. number of neighbors. The neighborhoods were computed using 1) co-occurrence-based metric D (solid line), and 2) context-based metric $Q^{H=1}$ (dash–dotted line). The reference nouns were taken from the RG dataset. (b) Percentage of neighbors that do not co-occur with the reference nouns vs. number of neighbors. In total, $1,000$ reference nouns were randomly selected from the lexicon. The neighborhoods were computed by the context-based metric Q^H. The percentage is shown for different values of H.

Given the importance of senses for the computation of semantic similarity, we attempt to quantify the performance of co-occurrence and context-based metrics with respect to the discovery of senses through their neighborhoods. The percentage of synonyms of reference nouns (taken from the RG dataset) that are included in the neighborhoods are presented in Fig.5(a) as a function of the neighborhood size. The sets of synonyms for each reference noun were created by consulting the WordNet synsets. The semantic neighborhoods were computed using either the co-occurrence metric D, or the context metric $Q^{H=1}$. In general, more synonyms are captured by the D metric compared to the $Q^{H=1}$ metric. This distinction is greater for neighborhoods that include more than 50 members.

Moreover, we investigate the effect of the context window H with respect to the selection of neighbors that do not co-occur with the reference nouns. The percentage of such neighbors computed by Q^H is depicted in Fig.5(b) for several sizes of the neighborhoods, and for four values of the contextual window size H. The percentages were computed for $1,000$ nouns that were randomly selected from the network. The best results are consistently obtained when using immediate context, i.e., $H = 1$, which can be attributed to the best performance of this window value for the case of context-based similarity computation (Iosif and Potamianos, 2010). This is also shown here in Table 1.

7.3.2 Neighborhood-based Metrics

The computation of semantic similarity consists of two basic steps: 1) computation of semantic neighborhoods, and 2) computation of similarity scores (the S metric in (8) and (9)), allowing for the following combinations.

Table 5. *Correlation for neighborhood-based metrics. Four combinations of the co-occurrence-based metric Dice (D) and the context-based metric $Q^{H=1}$ were used for the definition of semantic neighborhoods and the computation of similarity scores.*

Dataset	Neighbor selection	Similarity computation	Abbreviation for neighbor sel./ similarity comp.	Metrics		
				$M_{n=100}$	$R_{n=100}$	$E^{\theta=2}_{n=100}$
MC	co-occur.	co-occur.	(CC/CC)	0.90	0.72	**0.90**
MC	co-occur.	context	(CC/CT)	**0.91**	0.28	0.46
MC	context	co-occur.	(CT/CC)	0.52	**0.78**	0.56
MC	context	context	(CT/CT)	0.51	0.77	0.29
RG	co-occur.	co-occur.	(CC/CC)	**0.87**	0.67	**0.86**
RG	co-occur.	context	(CC/CT)	0.86	0.32	0.53
RG	context	co-occur.	(CT/CC)	0.58	**0.72**	0.61
RG	context	context	(CT/CT)	0.57	0.69	0.33
WS353	co-occur.	co-occur.	(CC/CC)	**0.64**	0.50	**0.64**
WS353	co-occur.	context	(CC/CT)	**0.64**	0.14	0.20
WS353	context	co-occur.	(CT/CC)	0.47	0.56	0.48
WS353	context	context	(CT/CT)	0.46	**0.57**	0.11

- Compute neighborhoods and similarity scores using a co-occurrence-based metric (CC/CC).

- Compute neighborhoods using a co-occurrence-based metric; compute similarity scores using a context-based metric (CC/CT).
- Compute neighborhoods using a context-based metric; compute similarity scores using a co-occurrence-based metric (CT/CC).
- Compute neighborhoods and similarity scores using a context-based metric (CT/CT).

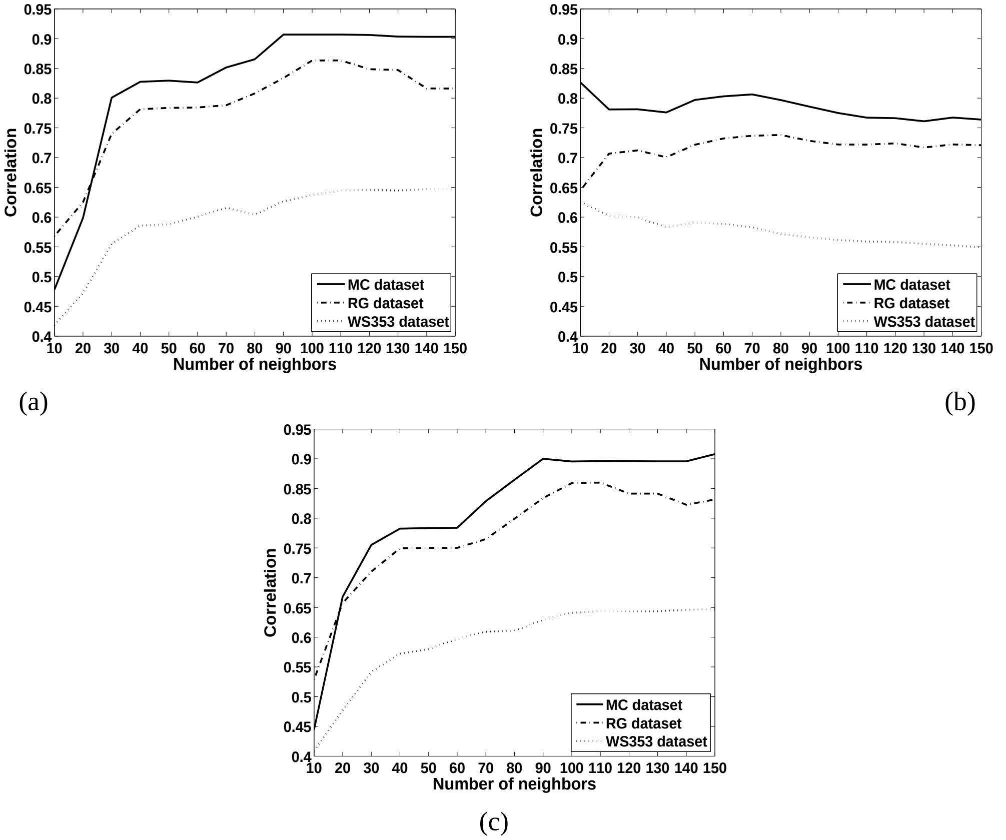
For the above approaches, the co-occurrence-based metric[15] D and the context-based metric $Q^{H=1}$ were used. The correlation results for the neighborhood-based metrics $M_{n=100}$, $R_{n=100}$, and $E^{\theta=2}_{n=100}$ for neighborhood size of 100 are presented in Table 5 (see the next paragraph for the choice of n). The use of a co-occurrence metric for neighbor selection achieves the highest results for all datasets, for $M_{n=100}$ and $E^{\theta=2}_{n=100}$, while, the context-based metric appears to be better for selecting neighbors for the correlation-based neighborhood metric $R_{n=100}$. The choice of the semantic similarity metric is of secondary importance for the $M_{n=100}$ and $R_{n=100}$ metrics, provided that the appropriate metric is used for neighborhood creation. For the $E^{\theta=2}_{n=100}$ metric however, only the (CC/CC) combination performs well. The results are significantly higher compared to the context-based baselines (see Table 1). The best $M_{n=100}$ and $E_{n=100}$ metrics also outperform the metrics that rely on web or corpus counts. Overall, utilizing network neighborhoods for estimating semantic similarity can achieve very good performance, and the type of metric (feature) used to select the neighborhood is a key performance factor.

Next, we investigate the performance of the metrics as a function of neighborhood size n. The performance of the M_n metric using co-occurrence-based metric D for neighbor selection, and $Q^{H=1}$ for similarity computation is shown in Fig.6(a). We observe that performance increases with n peaking around $n = 80-100$. The performance remains high also for $n > 100$. The performance of the R_n metric using $Q^{H=1}$ for neighbor selection and D for similarity computation is shown in Fig.6(b). The performance of R_n is relatively flat as a function of neighborhood size, achieving good performance even for small neighborhoods. The performance of the E^{θ}_n metric using D for both neighborhood selection and similarity estimation is shown in Fig.6(c). M_n and E^{θ}_n exhibit comparable performance, while both appear to be better than R_n for high values of n.

The correlation scores for the best performing neighborhood metrics ($M_{n=100}$, $R_{n=100}$ and $E^{\theta=2}_{n=100}$ for the (CC/CT), (CT/CC) and (CC/CC) approaches, respectively) are presented in Table 6 as a function of the number of snippets downloaded for each word in the network. The performance of the corresponding baseline metrics are also shown in Table 6, i.e., the D metric relying on corpus counts, and $Q^{H=1}$. We observe that the neighborhood metrics outperform the baseline performance for all datasets. All three neighborhood metrics consistently obtain better correlation performance as the number of snippets increases. Unlike neighborhood metrics, the performance of baseline metrics is not shown to improve as the number of snippets increases and plateaus around 300–500 snippets.

Next, we investigate the performance of the neighborhood metrics with respect to the number of concepts (nouns) included in the network. The concepts were randomly se-

[15] D achieved slightly higher performance than other co-occurrence metrics (not shown here for the sake of space).



(a)

(b)

(c)

Fig. 6. Performance vs. number of neighbors for neighborhood-based metrics: (a) maximum similarity of neighborhoods M_n: (CC/CT), (b) correlation of neighborhood similarities R_n: (CT/CC), and (c) sum of squared neighborhood similarities E_n^{θ}: (CC/CC).

lected; results are presented in Table 7 in the form of average correlation computed over ten runs. We experimented with various network sizes varying from 9 (0.1% of network) up to $8,752$ (100% of network) words. Regarding $M_{n=100}$ and $E_{n=100}^{\theta=2}$ metrics, performance improves as the network grows with best results around 4–5K words. Conversely $R_{n=100}$ perform best for small networks[16].

[16] Note that since the neighborhood size is set to be (up to) $n = 100$ for all experiments for the first two rows (with network size of 9 or 88 words) all available words in the network are used to construct the neighborhoods, i.e., the set of neighbors is the same for all words considered. The superior performance of $R_{n=100}$ for small network size is a strong indication that using a common set of words to compare semantic similarities on, works better than using each word's semantic neighbor. The approach of using a common set of "seed words" has been successfully applied to affective text analysis (Turney and Littman, 2002; Malandrakis et al., 2011) and warrants further research also for semantic similarity computation.

Table 6. *Performance with respect to the number of corpus snippets per noun for the baseline and the neighborhood-based metrics.*

Metric	Neighbor selection	Similarity computation	Dataset	Number of snippets per noun				
				50	100	200	500	1,000
Baseline	*not applicable*	co-occur. (corpus-based)	MC	0.24	0.31	0.43	0.57	**0.59**
			RG	0.35	0.42	0.56	**0.62**	0.60
			WS353	0.26	0.26	**0.27**	**0.27**	0.22
Baseline	*not applicable*	context	MC	0.35	0.52	**0.57**	0.54	0.53
			RG	0.38	0.45	0.50	**0.55**	0.52
			WS353	0.30	0.33	**0.34**	0.32	0.30
$M_{n=100}$	co-occur.	context	MC	0.54	0.61	0.71	0.88	**0.91**
			RG	0.54	0.60	0.73	0.83	**0.86**
			WS353	0.53	0.54	0.56	0.62	**0.64**
$R_{n=100}$	context	co-occur.	MC	0.28	0.49	0.67	0.73	**0.78**
			RG	0.42	0.60	0.68	0.69	**0.72**
			WS353	0.50	0.48	0.54	0.55	**0.56**
$E_{n=100}^{\theta=2}$	co-occur.	co-occur.	MC	0.56	0.61	0.69	0.83	**0.90**
			RG	0.57	0.61	0.72	0.81	**0.86**
			WS353	0.53	0.54	0.57	0.61	**0.64**

Table 7. *Performance of the neighborhood metrics for various network sizes.*

Num. of concepts in net.	Metrics								
	$M_{n=100}$			$R_{n=100}$			$E_{n=100}^{\theta=2}$		
	MC	RG	WS353	MC	RG	WS353	MC	RG	WS353
9	0.68	0.63	0.55	0.87	0.70	**0.60**	0.75	0.42	**0.66**
88	0.68	0.63	0.55	**0.88**	**0.79**	**0.60**	0.70	0.45	0.61
176	0.68	0.63	0.54	0.86	0.78	**0.60**	0.70	0.44	0.60
876	0.68	0.69	0.58	0.83	0.74	0.59	0.73	0.60	0.62
1,751	0.75	0.73	0.62	0.80	0.71	0.58	0.80	0.66	0.64
4,376	**0.95**	0.82	**0.68**	0.78	0.70	0.57	**0.95**	0.75	**0.66**
6,127	0.91	**0.86**	0.65	0.77	0.72	0.57	0.90	0.72	0.64
8,752	0.91	**0.86**	0.64	0.78	0.72	0.56	0.90	**0.86**	0.64

7.4 Fusion of Neighborhood Metrics

Next, we investigate the fusion of the best performing neighborhood metrics, M_n, R_n, and $E_n^{\theta=2}$, using the (CC/CT), (CT/CC), and (CC/CC) combinations, respectively (see Table 5). The fusion was performed as a weighted linear combination of their respective similarity scores. The largest dataset, i.e., WS353, was used for learning the weights of similarities using 10–fold cross validation. Then, the weights learned on (all of) WS353 were applied to the CM and RG datasets. Three different algorithms implemented in Weka[17] were applied for learning the weights, namely, linear regression, regression using Support Vector Machines (SVM), and regression trees. The performance of the fusion of metrics is presented in Table 8 for $n = 100$, along with the performance of the best individual neighborhood metric[18].

Table 8. *Performance for the fusion of neighborhood metrics.*

Metric/	Dataset		
Fusion algorithm	MC	RG	WS353
Best individual neighborhood metric	0.91	**0.86**	0.64
Linear regression	0.91	**0.86**	0.65
Regression using SVM	0.91	**0.86**	0.65
Regression trees	**0.94**	0.82	**0.73**

We observe that the performance of fusion using linear and SVM–based regression is almost identical to the performance of the best individual neighborhood metric. Performance gains are obtained using regression trees for the CM (from 0.91 to 0.94) and WS353 dataset (from 0.64 to 0.73). However, performance is worse on the RG dataset. This trend is probably due to the different distribution of the similarity scores in the datasets (MC dataset for example contains only highly similar or dissimilar word pairs, while RG contains more uniformly distributed similarity scores).

7.5 Comparison with Other Approaches

A comparison between our best results[19] and the performance of other similarity metrics is summarized in Table 9. The primary criterion for the selection of the presented metrics is the type of the exploited resources and corpora. This enables the comparison of knowledge– and data–driven approaches, while the latter often are the only feasible choice

[17] `http://www.cs.waikato.ac.nz/ml/weka/`

[18] We observed that the fusion algorithms exhibited similar (relative) performance for also other values of n (not reported here).

[19] As mentioned in Section 6 regarding the RG and WS353 datasets, we used their respective subsets covered by SemCor3. The same subsets were also used for the evaluation of the WordNet-based metrics.

for under-resourced languages. The approaches that are presented in Table 9 can be distinguished into two main categories: (i) use of knowledge resources, such as WordNet, (ii) use of large corpora, e.g., Wikipedia and corpora harvested from the web. In addition, we consider a third category dealing with the integration of (i) and (ii) within a machine learning-based framework.

Table 9. *Performance of several metrics/systems.*

Metric /	Resources /		Dataset		
System[a]	Corpora	ML[b]	MC	RG	WS353
Wup	WordNet	no	0.76	0.78	0.34
Res	WordNet + SemCor	no	0.77	0.80	0.37
Vector	WordNet + SemCor	no	0.85	0.79	0.47
WikiRelate!	Wikipedia	no	0.45	0.53	0.48
AAHKPS1	4 billion web docs	no	0.88	0.89	0.66
TypeDM	ukWaC + Wikipedia + BNC	no	–	0.82	–
IP	$28,000$ web docs: AND queries	no	0.88	–	–
IP$_s$	web doc snippets: AND queries	no	0.80	0.81	0.57
AAHKPS2	WordNet + 4 billions web docs	yes	0.92	0.96	0.78
SSS	WordNet + 9 million web doc snippets	yes	0.88	–	–
Proposed	$\sim$ 9 million web doc snippets: IND queries				
($M_{n=100}$)		no	0.91	0.87	0.64
($E^{\theta=2}_{n=100}$)		no	0.91	0.86	0.64
(Fusion)		yes	0.94	0.82	0.73

[a] The metrics/systems shown in full uppercase, e.g. IP, were abbreviated using the first letter of authors' last names.
[b] Use of machine learning.

Three basic types of WordNet-based metrics are included in category (i): path length-based (Wup), information content-based (Res), and metrics that exploit the synset glosses (Vector). Wup (Wu and Palmer, 1994) is a purely taxonomic metric based on the notion of the least common subsumer (LCS), i.e., the most specific concept that is the parent node of two words. The similarity between two words, w_i and w_j, is estimated as the depth (distance from root node) of their LCS, normalized by their individual depths (Pedersen, 2010). Wup is extended by the Res metric (Resnik, 1995) according to which the similarity of w_i and w_j is estimated as $Res(w_i, w_j) = -\log P(LCS(w_i, w_j))$, where $P(LCS(w_i, w_j))$ is the probability of the LCS of w_i and w_j estimated over a sense-tagged corpus (Pedersen, 2010). The lexical information that is included in the WordNet glosses is utilized by the Vector metric (Patwardhan and Pedersen, 2006) for the construction of co-occurrence vectors extracted from a sense-tagged corpus. The similarity between w_i and w_j is estimated as the similarity of their respective vectors. In this work, we applied the aforementioned WordNet-based metrics using the WordNet::Similarity module[20] , which incorporates the SemCor corpus (Pedersen and Michelizzi, 2004). More specifically, the similarity between two words was estimated according to (11) following the maximum sense similarity assumption (Resnik, 1995; Budanitsky and Hirst, 2006). Regarding category (ii), the WikiRelate! system (Strube and Ponzetto, 2006) includes various taxonomy-based metrics that are typically applied to the WordNet hierarchy. The basic idea behind WikiRelate! is to adapt these metrics to a hierarchy extracted from the links between the pages of the English Wikipedia. A very large corpus is exploited by AAHKPS1 consisting of four billion web documents that were acquired via crawling (Agirre et al., 2009). For the computation of semantic similarity several variations of structured and unstructured DSM were applied. An example of structured DSM is the TypeDM model (Baroni and Lenci, 2010), where a number of lexico-syntactic patterns were extracted from the concatenation of three different corpora, namely, the web-harvested ukWaC corpus[21] , the dump of the English Wikipedia, and the British National Corpus (BNC). Our previous work, IP, is an example of corpus creation using a relatively small number of web documents (Iosif and Potamianos, 2010). The basic idea was the use of conjunctive AND queries in order to retrieve documents in which the pair words co-occur. Also, we have replicated [22] our previous work using snippets instead of entire web documents (IP$_s$).

The third category that appears in Table 9 includes the following machine learning-based metrics/systems: AAHKPS2 and SSS. The basic approach behind AAHKPS2 (Agirre et al., 2009) is the use of regression in order to combine similarity scores that were computed using different resources and corpora. A corpus of four billion web documents was exploited and results were derived using 10-fold cross validation. A different approach was followed by the SSS system (Spanakis et al., 2009) according to which the WordNet was exploited in order to create thousands of word pairs denoting relations such as synonymy, meronymy, etc. These pairs were used for the formulation of web queries in order to create a corpus of snippets from which numerous lexico-syntactic patterns were extracted. The word similarity was estimated by a regression model considering the pattern frequencies as training features. The WS353 dataset was used for training excluding the pairs of the MC dataset, which were used for testing.

As it was expected, the exploitation of knowledge resources leads to high performance. The superiority of the Vector metric over the other WordNet-based metrics constitutes a successful paradigm regarding the exploitation of contextual features given that the word senses are knwon. The performance of the DSM-based approaches, i.e., AAHKPS1, TypeDM, and IP, is higher compared to the WordNet metrics. This observation is more interesting regarding the case of IP, where a relatively small corpus of web documents is used. The proposed metrics clearly outperform the resourced-based approaches: WordNet-

[20] `http://search.cpan.org/dist/WordNet-Similarity/`
[21] `http://wacky.sslmit.unibo.it/`
[22] As in IP, the top $1,000$ search results were acquired for each pair.

based (Wup, Res, Vector) and Wikipedia-based (WikiRelate!), and also obtain higher results than the paradigm of structured DSM (TypeDM). Also, the utilization of IND queries yields better performance compared to our previous work using AND queries (IP,IPs). Regarding unsupervised approaches (i.e., no use of machine learning), AAHKPS1 appears to be the closest competitor to the proposed metrics. Overall, the best performance is obtained by the machine learning-based approaches AAHKPS2 for the RG and WS353 datasets, and the fusion of $M_{n=100}$, $R_{n=100}$, and $E^{\theta=2}_{n=100}$ for the MC dataset. However, we believe that further validation is needed for the machine learning approaches given the limited size of the datasets and the dangers of overfitting. Overall, the proposed $M_{n=100}$ and $E^{\theta=2}_{n=100}$ metrics can be regarded among the best-performing unsupervised data-driven metrics, built upon an efficient and scalable approach for corpus creation using web data.

8 Conclusions

We have investigated the estimation of semantic similarity using semantic networks, following an unsupervised corpus-based approach. We have shown that it is possible to achieve state-of-the-art performance by encoding corpus statistics into a semantic network and then using the notion of semantic neighborhood to define novel semantic similarity metrics. The maximum neighborhood similarity metric performed the best when the semantic neighborhood was defined using co-occurrence metrics. We have also shown experimentally the importance of sense coverage and the validity of the maximum sense similarity assumption for context-based similarity metrics.

The fact that co-occurrence proved to be a good feature for selecting neighbors for the maximum similarity metric implies that co-occurrence is a good feature for sense discovery. Moreover, we have studied the effect of word proximity for the estimation of semantic similarity, showing that very good performance is obtained when words co-occur at sentential level. The success of context-based similarity for neighborhood selection for the correlation metric implies that context is a good feature for discovering attributes in a network. In addition, the use of a corpus in which the not so common words are well-represented and a large lexicon creates an informative corpus that efficiently encodes the semantics of polysemous words and leads to good performance. More research and experimentation is needed to verify these claims. Overall, the achieved results are amongst the highest reported in the literature for unsupervised corpus-based metrics. Last but not least, the proposed approach is efficient, scalable and requires linear web query complexity with respect to the lexicon size.

A series of preliminary experiments were conducted in order to investigate the applicability of the proposed approach to other semantic tasks, namely: compositional similarity estimation for noun-noun compounds, and word semantic similarity using networks of words and images. Very encouraging results were obtained for both tasks shown the extensibility of network-based DSM to compositional and multimodal semantic similarity tasks. Future work deals with the incorporation of network features, such as centrality measurements, for the creation of semantic neighborhoods. Further research is needed with larger multilingual networks to verify the universality of the proposed metrics.

9 Acknowledgements

Elias Iosif was partially funded by the Basic Research Programme, Technical University of Crete, Project Number 99637: “Unsupervised Semantic Relationship Acquisition by Humans and Machines: Application to Automatic Ontology Creation”. The authors wish to thank Nikos Malandrakis for many useful discussions, as well as, applying neighborhood–based similarity metrics to affective text computation.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–27.

Agirre, E. and Edmonds, P., editors (2007). *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Two graph-based algorithms for state-of-the-art WSD. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 585–593.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proc. Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *Proc. of International Conference on World Wide Web*, pages 757–766.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proc. of the Seventh International Conference on World Wide Web*, pages 107–117.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32:13–47.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proc. of the annual meeting of the Association for Computational Linguistics: HLT*, pages 120–126.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Erk, K. and Padó, S. (2010). Exemplar-based models for word meaning in context. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Ferrer-I-Cancho, R. and Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London, Series B, Biological Sciences*, 268:2261–2266.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Gracia, J., Trillo, R., Espinoza, M., and Mena, E. (2006). Querying the web: A multiontology disambiguation method. In *Proc. of International Conference on Web Engineering*, pages 241–248.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Harrington, B. (2010). A semantic network approach to measuring relatedness. In *Proc. of the 23rd International Conference on Computational Linguistics*, pages 356–364.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Haveliwala, T., Gionis, A., Klein, D., and Indyk, P. (2002). Evaluating strategies for similarity search on the web. In *Proc. of the 11th International World Wide Web Conference*, pages 432–442.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of Conference on Computational Linguistics*, pages 539–545.

Hughes, T. and Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 581–589.

Ide, N. and Véronis, J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Iosif, E. and Potamianos, A. (2010). Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1637–1647.

Iosif, E. and Potamianos, A. (2012a). Minimum error semantic similarity using text corpora constructed from web queries. *IEEE Transactions on Knowledge and Data Engineering* (submitted to).

Iosif, E. and Potamianos, A. (2012b). Semsim: Resources for normalized semantic similarity computation using lexical networks. In *Proc. Eighth International Conference on Language Resources and Evaluation*, pages 3499–3504.

Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of International Conference on Research on Computational Linguistics*, pages 19–33.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification in WordNet. In Fellbaum, C., editor, *An Electronic Lexical Database*, pages 265–283. MIT Press.

Lemaire, B. and Denhière, G. (2004). Incremental construction of an associative network from a corpus. In *Proc. of the 26th Annual Meeting of the Cognitive Science Society*, pages 825–830.

Malandrakis, N., Iosif, E., and Potamianos, A. (2012). DeepPurple: Estimating sentence semantic similarity using n-gram regression models and web snippets. In *Proc. of the First Joint Conference on Lexical and Computational Semantics*, pages 565–570.

Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2011). Kernel models for affective lexicon creation. In *Proc. Interspeech*, pages 2977–2980.

Meng, H. and Siu, K.-C. (2002). Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):172–181.

Mihalcea, R. and Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Miller, G. and Charles, W. (1998). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R. and Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 116–126.

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

Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proc. of the EACL Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8.

Pedersen, T. (2010). Information content measures of semantic similarity perform better without sense-tagged text. In *Proc. of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 329–332.

Pedersen, T. and Michelizzi, S. P. J. (2004). Wordnet::similarity - measuring the relatedness of concepts. In *Proc. of AAAI*, pages 1024–1025.

Radev, D. and Mihalcea, R. (2008). Networks and natural language processing. *AI Magazine*, 29(3):116–126.

Reddy, S., Klapaftis, I., McCarthy, D., and Manandhar, S. (2011). Dynamic and static prototype vectors for semantic composition. In *Proc. of the 5th International Joint Conference on Natural Language Processing*, pages 705–713.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxanomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pages 448–453.

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

Sebastiani, F. and Ricerche, C. N. D. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Spanakis, G., Siolas, G., and Stafylopatis, A. (2009). A hybrid web-based measure for computing semantic relatedness between words. In *Proc. of the 21st International Conference on Tools with Artificial Intelligence*, pages 441–448.

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Proc. of 21st National Conference on Artificial intelligence*, pages 1419–1424.

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, P. and Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.

Turney, P. D. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. of the European Conference on Machine Learning*, pages 491–502.

Véronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language*, 18(3):223–252.

Vitanyi, P. (2005). Universal similarity. In *Proc. of Information Theory Workshop on Coding and Complexity*, pages 238–243.

Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 1093–1099.

Wojtinnek, P.-R., Pulman, S., and Völker, J. (2012). Building semantic networks from plain text and wikipedia with application to semantic relatedness and noun compound paraphrasing. *International Journal of Semantic Computing (IJSC). Special Issue on Semantic Knowledge Representation.*, 6(1):67–91.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proc. of the annual meeting on Association for Computational Linguistics*, pages 133–138.

Zipf, G. K. (1965). *The Psycho-Biology of Language*. MIT Press.