Automatic Extraction of Semantic Relations by Using Web Statistical Information*

Valeria Borzì[†], Simone Faro[†,*], Arianna Pavone[‡]

[†]Dipartimento di Matematica e Informatica, Università di Catania
Viale Andrea Doria 6, I-95125 Catania, Italy

[‡]Dipartimento di Scienze Umanistiche, Università di Catania
Piazza Dante 32, I-95124 Catania, Italy

[*]faro@dmi.unict.it

Abstract. A semantic network is a graph which represents semantic relations between concepts, used in a lot of fields as a form of knowledge representation. This paper describes an automatic approach to identify semantic relations between concepts by using statistical information extracted from the Web. We automatically constructed an associative network starting from a lexicon. Moreover we applied these measures to the ESL semantic similarity test proving that our model is suitable for representing semantic correlations between terms obtaining an accuracy which is comparable with the state of the art.

1 Introduction

In recent years, with the increase of the information society, lexical knowledge, i.e. all the information that is known about words and all the relationships among them, is becoming a core research topic in order to understand and categorize all subjects of interest [14]. We need lexical knowledge to know how words are used in different ways to express different meanings [13].

An *associative network* is a labeled directed (or undirected) graph representing relational knowledge. Each vertex of the graph represents a concept and each edge (or link) represents a relation between concepts. Such structures are used to implement cognitive models representing key features of human memory.

Specifically, when two concepts, x and y, are thought simultaneously, they may become linked in memory. Subsequently, when one thinks about x, then y is likely to come to mind as well. Thus multiple links to a concept in memory make it easier to be retrieved because of many alternative routes to locate it.

A *semantic network* is an associative network where we introduce labels on the links between words [3], [14]. Labels represent the kind of relation between the two given concepts, such as "is-a", "part-of", "similar-to" and "related-to".

* This work has been supported by project PRISMA PON04a2 A/F funded by the Italian Ministry of University and Research within the PON 2007-2013 framework.

Aristotle firstly described some of the principles governing the role of associative networks and categories in memory, while the concept of semantic network dates back to the 3rd century AD when the greek philosopher Porphyry, in his commentary on Aristotle's categories, drawn the oldest known semantic network, called *Porphyry's tree*. Despite its age, the Tree of Porphyry represents the common core of all modern type concept hierarchies.

The potential usefulness of large scale lexical knowledge networks can be attested by the number of projects and the amount of resources that have been dedicated to their construction [3], [14]. Creating such resources manually is a difficult task and it has to be repeated from ex novo for each new language. However there are a lot of important resources of this kind. Among the others the most relevant are WordNet, Wikipedia and BabelNet.

WordNet [3], is a lexical knowledge resource. It is a computational lexicon of the English language based on psycholinguistic principles. A concept in WordNet is represented as a synonym set (called *synset*), i.e. the set of words that share the same meaning. Synsets are related to each other by means of many lexical and semantic relations. Wikipedia instead is a multilingual Web-based encyclopedia. It is a collaborative open source medium edited by volunteers to provide a very large wide-coverage repository of encyclopedic knowledge. The text in Wikipedia is partially structured, various relations exist between the pages themselves. These include redirect pages (used to model synonymy), disambiguation pages (used to model homonymy and polysemy), internal links (used to model relations between terms) and categories. Finally, BabelNet [14] is a multilingual encyclopedic dictionary and a semantic network, currently covering 50 languages, created by linking Wikipedia network to WordNet, thus it includes lemmas which denote both lexicographic meanings and encyclopedic ones. However, a widely acknowledged problem with the above semantic networks is that they implements links which represent uniform distances between terms, while conceptual distances in real world relations between concepts could have a wide variability. As a consequence we find in Wikipedia, or in BabelNet, links between very close concepts but also links between terms that are conceptually distant, and no measure leading to distinguish between them.

In this paper we describe an automatic approach to identify semantic relatedness between concepts, by using statistical information extracted from the Web. We then use such semantic measure to construct a weighted associative network starting from the English WordNet lexicon, augmented with Wikipedia encyclopedic entities. From our preliminary experimental results it turns out that our presented approach can be efficiently used to identify semantic relatedness between concepts. The paper is organized as follows. In Section 2 we introduce the concept of semantic relatedness, which is particularly connected with our results, and present the most significant results on computing such measures. Then in Section 3 we introduce a new model The construction process is described in Section 4. In the next sections we present some experimental results (Section 5) and some examples (Section 6) in order to evaluate the effectiveness of the new presented model. We discuss our results and describe future works in Section 7.

2 Measuring Semantic Relatedness

Lexical semantic relatedness is a measure of how much two terms, words, or their senses, are semantically related. It has been well studied and categorized in linguistics.

Evaluating semantic relatedness using network representations is a problem with a long history in artificial intelligence and psychology, dating back to the spreading activation approach of Quillian [15] and Collins and Loftus [2]. It is important for many natural language processing or information retrieval applications. For instance, it has been used for spelling correction, word sense disambiguation, or coreference resolution. It has also been shown to help inducing information extraction patterns, performing semantic indexing for information retrieval, or assessing topic coherence.

Semantic relations between terms include typical relations such as

- synonymy: identity of senses as "automobile" and "car";
- antonymy: opposition of senses such as "fast" and "slow";
- hypernymy or hyponymy: such as "vehicle" and "car";
- meronymy or holonymy: part-whole relation such as "windshield" and "car".

Most of the recent research has focused on *semantic similarity* [17], [21, 22], [6], [18], which represents a special case of semantic relatedness. For instance, antonyms are related, but not similar. Or, following Resnik [17], "car" and "bicycle" are more similar (as hyponyms of "vehicle") than "car" and "gasoline", though the latter pair may seem more related in the world.

Thus, while typical relations implying sense similarity are widely represented in lexicons like WordNet [3] and BabelNet [14], the latter types of relations are usually not always included in state-of-the-art ontologies, although they are relevant in conceptual connections between terms. Such relations include, for instance, the following

- synecdoche: a portion of something refers to the whole, as "information" for a "book" or as "cold" for the "winter";
- antonomasia: an epithet for a proper name, as "The Big Apple" for "New York" or as "The Conqueror" for "Caesar";
- trope: a figurative meaning for its literal use, "to bark" for "to shout".

Current approaches to address semantic relatedness can be categorized into three main categories: lexicon-based methods, corpus-based methods, and hybrid approaches.

In a *lexicon-based methods* the structure of a lexicon is used to measure semantic relatedness. Such approach consists in evaluating the distance between the nodes corresponding to the terms being compared: the shorter the path from one node to another, the more similar they are. Such approaches rely on the structure of the lexicon, such as the semantic shortest link path [11], the depth of the terms in the lexicon tree [23], the lexical chains between synsets and their relations [6], or on the type of the semantic edges [21]. Finally in [18] the authors

use all 26 semantic relations found in WordNet in addition to information found in glosses to create an explicit semantic network.

However, a widely acknowledged problem with this approach is that it relies on the notion that links in the taxonomy represent uniform distances [18]. Unfortunately, this is difficult to define, much less to control. In real lexicons, there could be wide variability in the distance covered by a single relation link.

Differently, *corpus-based methods* use statistical information about words distribution extracted from a large corpus to compute semantic relatedness. For instance in [19, 5] the authors used the statistical information from Wikipedia.

For the sake of completeness we mention also *hybrid methods* which use a combination of corpus-based and lexicon-based methods [7, 1] to compute semantic relatedness between two terms.

3 A New Model for Directional Semantic Relatedness

In this section we formalize the model of semantic relatedness which has been used to construct our associative network. Unlike state-of-the-art networks, in our structure the edges represent a certain correlation between two terms and give a measure of such relations. Thus we obtain a weighted network where terms closely related have small distances while weakly correlated terms have a great distance. The distance between two nodes of the network is inversely proportional to their semantic correlation which we measure by an *attraction coefficient*. The closer is the semantic correlation between the two words, the greater is their attraction. In turn the semantic attraction between two different terms is a function of their *usage coefficient*, i.e. a numeric value which measures how much the corresponding term is used in a given language.

In what follows we formalize this concepts and give the mathematical definitions of the formulas we use for computing the semantic relatedness in our network.

The usage coefficient

All natural languages like English consist of a small number of very common words, a larger number of intermediate ones, and then an indefinitely large set of very rare terms. We define the *usage coefficient* (U.C.) of a lexical term x, for a given language $\mathcal{L}$, as a value indicating how much x is used in $\mathcal{L}$. Such coefficient has been classically computed as the frequency of the term x in large corpora as the Oxford English Corpus[1], the Brown Corpus of Standard American English[2] or Wikipedia[3].

In order to give a real estimate of the frequency of a given term we compute the U.C. of words as a function of the number of pages resulting in a Google

[1] `http://www.oxforddictionaries.com`

[2] `http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html`

[3] `http://en.wikipedia.org`

search[4] for the term x. Specifically, for each term x contained in the English ontology, we performed a query on Google for x and use the number of page results for computing the U.C. of the term. We use the symbol $\rho(x)$ to indicate the U.C. of a term x.

Although the Google search engine does not guarantee the ability to return the exact number of results for any given search query[5] such value can be considered a good estimate of the actual number of results for the search request [16, 8]. We observed an upper bound for the number of page results retrieved by Google, i.e. MAX_RESULTS = 25.270 millions of results.

The U.C. of a term x is then computed by

$$\rho(x) = \frac{\text{PAGE\_RESULTS}(x)}{\text{MAX\_RESULTS}}$$

In our search we activate automatic filtering feature in order to reduces undesirable results such as duplicate entries. Moreover we filter search results by language and we use the `allintext:` operator[6] in order to reduce the search to internal text of the web pages.

The co-occurrence usage coefficient

Given a set of k terms, $\{x_1, x_2, \ldots, x_k\}$, of a given language $\mathcal{L}$, the *co-occurrence usage coefficient* (C.U.C.) of the terms x_i, is a value indicating how much such terms co-occur together in any context of the language. As before, we compute the C.U.C. as the number of pages resulting from a Google query for $x_1 \wedge x_2 \wedge \ldots \wedge x_k$, divided my the constant MAX_RESULTS.

We use the symbol $\rho(x_1 : x_2 : \ldots : x_k)$ to indicate the C.U.C. of the set $\{x_1, x_2, \ldots, x_k\}$.

By the definition given above it is trivial to observe that, for each $i = 1, \ldots, k$, the property $\rho(x_i) > \rho(x_1 : x_2 : \ldots : x_k)$ holds.

The attraction coefficient

A straightforward way to compute a similarity coefficient between two lexical terms is to use the *Jaccard similarity coefficient*, a statistic index introduced for comparing the similarity and diversity of sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. More formally if $\rho(x)$ and $\rho(y)$ are the U.C. of terms x and y, respectively, and $\rho(x : y)$ is their co-occurrence coefficient, the Jaccard similarity coefficient of x and y can be computed by using the following formula

$$jacc(x : y) = \frac{\rho(x : y)}{\rho(x) + \rho(y) - \rho(x : y)}$$

[4] `www.google.com`

[5] `http://www.google.com/support/enterprise/static/gsa/docs/admin/72/gsa_doc_set/xml_reference/appendices.html`

[6] `http://www.googleguide.com/advanced_operators_reference.html`

Such similarity coefficient has been used in [19], in combination with a lexicon-based approach, to measure the similarity relatedness of two terms. However it defines a symmetric semantic relation between x and y, thus assuming that $jacc(x : y) = jacc(y : x)$, which does not reflect the real world representation of associative networks where relations are, in general, represented by direct edges, i.e. the measure of the relation between x and y could be different from the measure of the relation between y and x.

Example 1. The terms "gasoline" and "car" are undoubtedly related in real world, thus if we think to gasoline the term "car" comes to mind with a great probability. However the contrary is not true, if we think to a car probably other terms come to mind with higher probability, like "road" or "parking". So we can say that "gasoline" is more related with "car" than viceversa.

In our model we define an unidirectional measure of semantic similarity between two terms. Specifically, the *attraction coefficient* (A.C.) of a lexical term x, on another term y of the same language, measures the semantic correlation of y towards x. In other words it is a numerical value evaluating how much the term x is conceptually related with term y (the contrary is not necessary).

More formally, let x and y two lexical terms, and let $\rho(x)$ and $\rho(y)$ the U.C. of x and y, respectively. Moreover let $\rho(x : y)$ be their co-occurrence coefficient. Then the attraction coefficient of y on x is defined by

$$\varphi(x \to y) = \frac{\rho(x : y)}{\rho(x)} \tag{1}$$

The following properties follow directly from the above definition and they are trivial to prove.

Property 1. If x and y are two lexical terms of $\mathcal{L}$, then the A.C. of y on x is a real number between 0 and 1. Formally

$$0 \leq \varphi(x \to y) \leq 1$$

Property 2. If x and y are two lexical terms of $\mathcal{L}$, and $\rho(x) > \rho(y)$ then the A.C. of x on y is greater than the A.C. of y on x. Formally

$$\rho(x) > \rho(y) \Longrightarrow \varphi(y \to x) > \varphi(x \to y)$$

Due to Property 2 it turns out that lexical terms with a huge U.C. are more attractive than other terms with smaller coefficient. This is the case, for instance, of general terms as "love", "man", "science", "music" and "book".

Example 2. Consider the numerical values related to the terms "bark", "kennel", "dog" and "man", presented in the following table, where the U.C. are expressed in million of results.

U.C.		C.U.C.		A.C.		
$\rho(\mathsf{bark})$	$\sim 0{,}043$	$\rho(\mathsf{bark} : \mathsf{dog})$	$\sim 0{,}012$	$\varphi(\mathsf{bark} \to \mathsf{dog})$	~ 0.28	(a)
$\rho(\mathsf{kennel})$	$\sim 0{,}026$	$\rho(\mathsf{bark} : \mathsf{man})$	$\sim 0{,}015$	$\varphi(\mathsf{dog} \to \mathsf{bark})$	~ 0.01	(b)
$\rho(\mathsf{dog})$	$\sim 0{,}813$	$\rho(\mathsf{kennel} : \mathsf{dog})$	$\sim 0{,}016$	$\varphi(\mathsf{bark} \to \mathsf{man})$	~ 0.35	(c)
$\rho(\mathsf{man})$	$\sim 1{,}000$	$\rho(\mathsf{kennel} : \mathsf{man})$	$\sim 0{,}005$	$\varphi(\mathsf{kennel} \to \mathsf{dog})$	~ 0.62	(d)
		$\rho(\mathsf{dog} : \mathsf{man})$	$\sim 0{,}372$	$\varphi(\mathsf{kennel} \to \mathsf{man})$	~ 0.19	(e)
				$\varphi(\mathsf{dog} \to \mathsf{man})$	~ 0.45	(f)

The term "bark" directly calls to mind the term "dog", since the bark is a prerogative of dogs, so that we can say that "bark" is semantically attracted by "dog" (a: 0.28). On the other hand, the contrary is not true since "dog" not necessarily calls to mind the term "bark" (b: 0.01), which is only one of the many inherent attitudes of a dog. Observe also that "bark" has a figurative meaning which can be applied to men, so it is semantically attracted also by the term "man" (d: 0.35). Differently the term "kennel" is strongly attracted by "dog" (d: 0.62) and is subject only to a feeble conceptual attraction by "man" (e: 0.19). The term "dog" is instead semantically attracted by the term "man" (f: 0.45), since the dog is the most popular domestic animal.

4 Building the Directed Semantic Graph.

We construct our semantic network starting from state of the art lexicon resources and by enriching them with new information and new semantic relations induced by the relatedness model described above. Specifically we start from the English WordNet semantic network.

The algorithm for building the corresponding directed semantic graph is depicted in Figure 1 and is named BUILDNETWORK. It takes as input the set $\mathcal{L}$ of all terms of the lexicon and constructs a directed weighted graph where each term x of the lexicon is a node in the graph, and directed links between two nodes represent semantic relations between the corresponding terms. Each link is associated with a weight value representing the attraction coefficient between the two related terms. The construction is divided in two steps, a bootstrap process and an exploration process, as described below.

The Bootstrap Process. In the bootstrap process (see Figure 1, on the left) the algorithm initializes the usage coefficient $\rho(x)$ for each term x of the set $\mathcal{L}$ (lines 2-3). In addition, for each term x, the algorithm also initializes the set, $\Psi(x)$ of all terms y such that $\varphi(x \to y) \geq \delta$, for a given bound δ (lines 4-12). In our construction we set $\delta = 0.1$. Specifically the set $\Psi(x)$ initially consists in all terms y which are related to x in the lexicon (line 5). In addition $\Psi(x)$ is augmented with the set of all significant terms from its definition, excluding all those words (conjunctions, adverbs, pronouns) that will not be particularly useful in the construction of the semantic field of x (line 6). Then all terms in the set $\Psi(x)$ are investigated in order to compute the attraction coefficient $\varphi(x \to y)$ (lines 7-10). During this process the algorithm deletes from $\Psi(x)$ all term y such that $\varphi(x \to y) < \delta$ (lines 11-12).

```
BOOTSTRAP(L)
 1. for each x ∈ L do
 2.     if ρ(x) = null do
 3.         ρ(x) ← getUC(x)
 4.     Ψ(x) ← ∅
 5.     Ψ(x) ← Ψ(x) ∪ getRelated(x)
 6.     Ψ(x) ← Ψ(x) ∪ getDefinition(x)
 7.     for each y ∈ Ψ(x) do
 8.         if ρ(y) ← null do
 9.             ρ(y) ← getUC(y)
10.         φ(x → y) ← ρ(x : y)/ρ(x)
11.         if (φ(x → y) < δ) then
12.             Ψ(x) ← Ψ(x) \ {y}
13.     explored(x) ← 0

EXPLORE(x)
 1. explored(x) ← 1
 2. for each y ∈ Ψ(x) do
 3.     if (explored(y) = 0) then
 4.         EXPLORE(y)
 5.     for each z ∈ Ψ(y) do
 6.         φ(x → z) ← ρ(x : z)/ρ(x)
 7.         if (φ(x → z) < δ) then
 8.             Ψ(x) ← Ψ(x) ∪ {z}

BUILDNETWORK(L)
 1. bootstrap(L)
 2. for each x ∈ L do
 3.     if (explored(x) = 0) then
 4.         EXPLORE(x)
```

Fig. 1. The algorithm which construct the semantic directed network. The construction makes use of two procedures, the BOOTSTRAP procedure and an EXPLORE procedure.

The Exploration Process. The next step of the algorithm consists in exploring each node graph by setting a recursive process (see Figure 1, on the right). For each term x, the flag explored(x) allows the algorithm to keep track of nodes already analyzed (a value set to 1), and nodes not yet explored (a 0 value). During the exploration process of the node x, the algorithm try to increase the set $\Psi(x)$ by adding new related terms contained in the lexicon. To do that the algorithm firstly recursively explore all neighbors y of node x (lines 3-4), i.e. all terms in the set $\Psi(x)$, and then it tries to add new links from x to all the neighbor nodes of y (lines 5-8). In other words, if the term x is semantically attracted by the term y and the latter is attracted by the term z, then the algorithm tries a possible relation between x and z. Observe that If a new node z enters the set $\Psi(x)$ (line 8) then all its neighbors will be considered for inclusion in the set. This process continues until all terms have been explored.

5 First Experimental Results

To test our approach to semantic relatedness between two terms of the lexicon, we evaluated it on a synonym identification test. Although different tests are available on the net, as for instance the WordSimilarity-353 similarity test[7], the one we experimented with is the larger English as a Second Language (ESL) test, which was first used by Peter Turney in [22] as an evaluation of algorithms measuring the degree of similarity between words. Specifically the ESL test includes 50 synonym questions. Each question includes a sentence, providing context for

[7] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`

the question, containing an initial word, and a set of options from which the most synonymous word must be selected. The following is an example question taken from ESL data set:

To [firmly] refuse means to never change your mind and accept
1. steadfastly
2. reluctantly
3. sadly
4. hopefully

The results are measured in terms of accuracy. For each question with initial word x and option words $\{y_1, y_2, y_3, y_4\}$ we compute the attraction coefficients $\varphi(y_i \to x)$, for $i = 1 \dots 4$ and put them in a decreasing order. Then we gave a decreasing score to each option word, from 4 to 1. Then the accuracy is computed as the sum of the scores obtained in the 50 questions compared with a full score result. The results of our approach, along with other approaches, on the 50 ESL questions are shown in Table 4. Our approach has achieved an accuracy of 84% on the ESL test, which is slightly better than the reported approaches in the literature. It should be noted that sometimes the difference between two approaches belonging to the same category are merely a difference in the data set used (Corpus or Lexicon) rather than a difference in the algorithms. Also, the ESL question set includes a sentence to give a context for the word, which some approaches (e.g. [22]) have used as an additional information source; we on the other hand, did not make use of the context information in our approach.

Approach	Year	Category	Accuracy
Resnik [17]	1995	Hybrid	32.66%
Leacock and Chodorow [11]	1998	Lexicon	36.00%
Lin [12]	1998	Hybrid	36.00%
Jiang and Conrath [10]	1997	Hybrid	36.00%
Hirst and St-Onge [6]	1998	Lexicon	62.00%
Turney [22]	2001	Corpus	74.00%
Terra and Clarke [20]	2003	Corpus	80.00%
Jarmasz and Szpakowicz [9]	2003	Lexicon	82.00%
Tsatsaronis et al. [21]	2010	Lexicon	82.00%
Siblini and Kosseim [18]	2013	Lexicon	84.00%
Our Approach	2014	Corpus	84.00%

Table 1. Results with the ESL Data Set.

6 Some Examples

In this section we present some experimental evidences related with the structure of the semantic net which has been constructed at the date of the paper

submission (January 16th, 2014). This is the reason why some terms are not depicted in the semantic nets, since they where not still added. In particular we briefly discuss portions of the semantic net connected with the terms "book" and "conquest". We present measures of relatedness between connected terms in both graphical and tabular forms. In Figure 2 and Figure 3 the diameter of a node representing a term x is proportional to its U.C. $\rho(x)$. Concentric circles represent distances from the main term, ranging from 1.0 (the innermost) to 0.3 (the outmost).

The network around "book". The term "book" has a very large semantic network and attracts different related words, since its U.C. is very large. We can observe that both terms "book" and "information" got the same U.C. value. Moreover their A.C. is equal to 1. This means that the two terms often occur together. Thus their relation can be interpreted as a synecdoche, which is distinguished by metonymy because it is based on quantitative relationships, through the broadening of meaning. Therefore it is assumed that "book" is a medium which convey "information" in general The terms "magazine", "title" and "cover" are positioned very close to the center and they are therefore strongly related to "book". Furthermore the relationships between the various terms of the semantic network are bidirectional. Thus, for example, the term "book" is strictly related to "cover" by a relationship of metonymy, viceversa the term "cover" is strictly related to "book" but with a relation of hyponymy, thus "book" is the hyperonymy term and "cover" is the hyponym. In other cases we notice that the semantic connections are unidirectional, for example the relation between "book" and "publishing house", where the latter term is directly related to the book, but the book is not directly related to "publishing house".

The network around "conquest". Table 3 shows the twenty closer lexical terms related with the term "conquest", while Figure 3 shows a graphical representation of the portion of the semantic network containing all terms related with "conquest". Typical relations of hyponymy and hypernyms can be found as "conquest" and "war", or "conquest" and "battle". A relation of metonymy can be read in the connection between "conquest" and "strategy".

Also, observing results shown in Table 3 we find a very interesting relation between "conquest" and "Caesar". Analyzing the results it is possible to notice that the two terms are strongly related, and also in this case we find a figure of speech, the antonomasia: the term "conquest" (as root of "conqueror") can be considered as representative of the term "Caesar", indeed. The relation between "conquest" and "attack" can be read as a metonymy, as cause-and-effect relation.

In addition, the relation between "conquest" and "freedom" can be read as a trope relation, since "conquest" here is used in its figurative meaning in place of "achievement". Similarly the same term is used, in connection with "love", with a figurative meaning in place of "seduction".

x	U.C.	C.U.C.	A.C. book	A.C. x	x	U.C.	C.U.C.	A.C. book	A.C. x
book	1,000								
information	1,000	1,000	1,00	1,00	copybook	0,001	0,000	0,52	-
magazine	0,895	0,830	0,93	0,83	ebook	0,192	0,091	0,47	-
cover	1,000	0,922	0,92	0,92	periodical	0,010	0,005	0,47	-
title	1,000	0,746	0,75	0,75	monograph	0,013	0,006	0,46	-
review	1,000	0,694	0,69	0,69	collection	1,000	0,444	0,44	0,44
school	1,000	0,624	0,62	0,62	press	1,000	0,421	0,42	0,42
publishing house	0,007	0,005	0,62	-	education	1,000	0,416	0,42	0,42
fiction	0,006	0,141	0,58	-	thriller	0,075	0,031	0,41	-
author	1,000	0,539	0,54	0,54	Gutenberg	0,008	0,003	0,37	-
word	1,000	0,535	0,54	0,54	reader	0,508	0,180	0,35	-

Table 2. The twenty lexical terms which have been found to be semantically closer to the term book. The number of results are expressed in millions of pages (Mr).

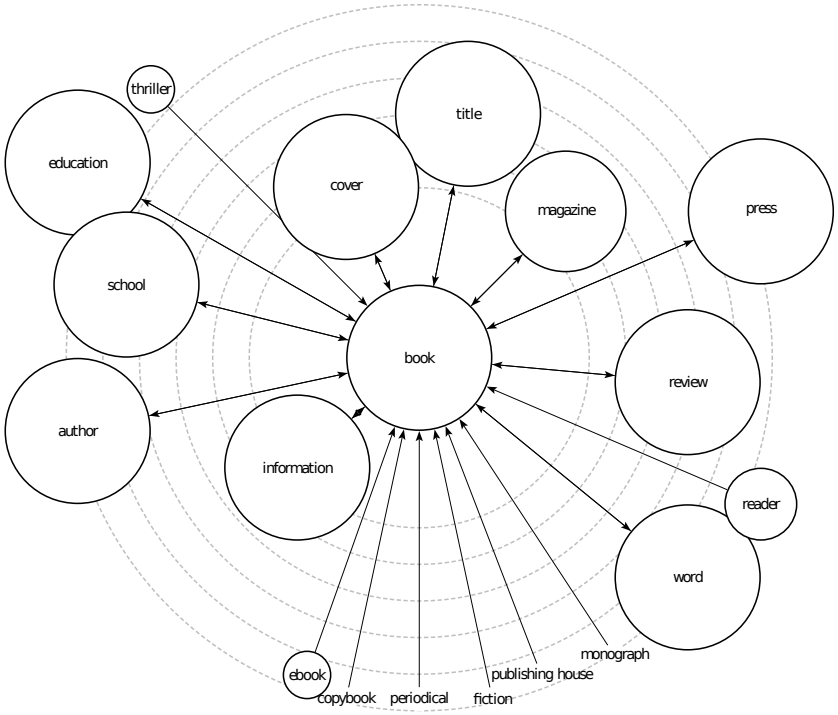

Fig. 2. A portion of the Semantic Net of the lexical term book. The diameter of a term x of the net if proportional to its U.C. $\rho(x)$. Concentric circles represent distances from the main term book.

x	U.C.	C.U.C.	A.C. conquest	A.C. x	x	U.C.	C.U.C.	A.C. conquest	A.C. x
conquest	0,029								
tyran	0,001	0,001	0,90	-	history	1,000	0,014	-	0,53
Athene	0,001	0,001	0,59	-	battle	0,579	0,009	-	0,30
military	0,026	0,009	0,34	-	right	1,000	0,014	-	0,49
Caesar	0,052	0,027	0,51	0,92	war	0,961	0,013	-	0,46
people	0,430	0,017	-	0,57	attack	0,497	0,007	-	0,24
empire	0,196	0,007	-	0,26	man	1,000	0,013	-	0,45
soldier	0,136	0,005	-	0,16	land	1,000	0,012	-	0,40
freedom	0,298	0,007	-	0,26	age	1,000	0,011	-	0,37
strategy	0,314	0,007	-	0,25	love	1,000	0,010	-	0,33
science	1,000	0,007	-	0,26	field	1,000	0,008	-	0,28

Table 3. The twenty lexical terms which have been found to be semantically closer to the term conquest. The number of results are expressed in millions of pages (Mr).

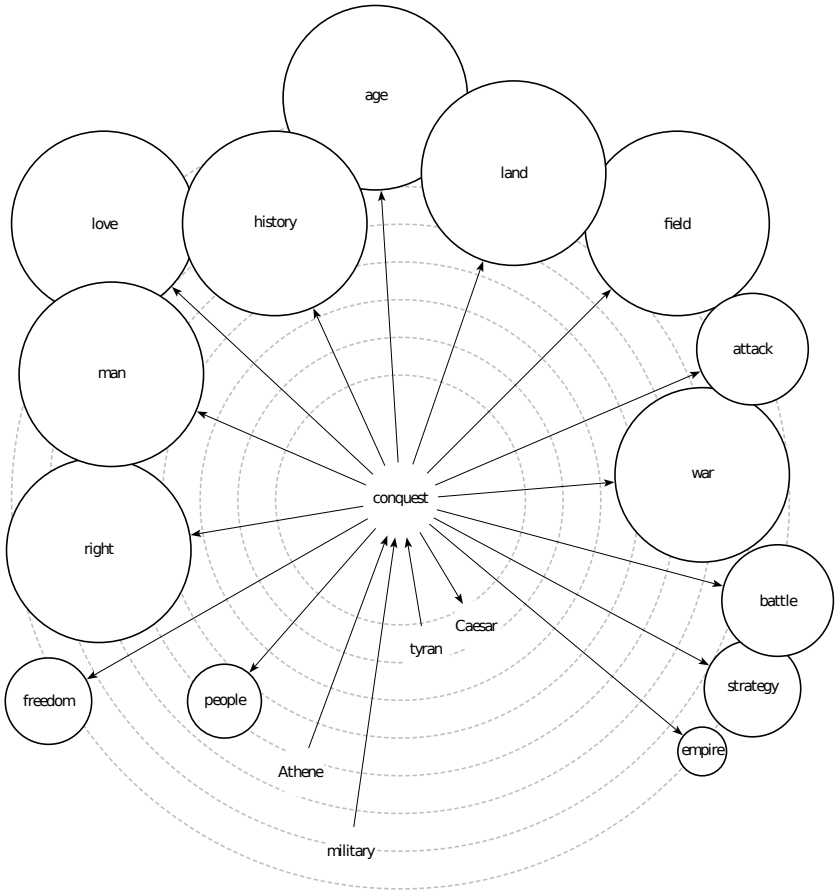

Fig. 3. A portion of the Semantic Net of the lexical term conquest. The diameter of a term x of the net if proportional to its U.C. $\rho(x)$. Concentric circles represent distances from the main term conquest.

7 Conclusions and Future Works

In this paper we described the construction of a semantic associative network for the English language. We start from the state-of-the-art semantic networks, as WordNet and Wikipedia, and enrich them with new informations measuring how much a term is used in practice. Then our algorithm explores the entire network in order to delete or add new semantic link according to a given model of directional semantic relatedness, based on statistical informations extracted from the Web. We then applied these measures to a real-world NLP task such as the ESL semantic similarity test. Our results show that our model is suitable for representing semantic correlations between terms obtaining an accuracy which is comparable with the state of the art.

Our algorithm is still exploring the network in order to complete the process of connecting all related terms. From our preliminary observations it turns out that several connections have been identified which do not appear in typical lexicons.

In future works we intend to construct a similar structure for the Italian language. Moreover we would like to perform additional experimental evaluation in order to test our model in field of semantic similarity or semantic relatedness.

Acknoledgements

We wish to thank Peter Turney for having provided the English as a Second Language (ESL) similarity test and for his precious suggestions.

References

1. Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa: A study on similarity and relatedness using distributional and wordnet-based approaches. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 19–27, Boulder, June (2009)
2. A. Collins and E. Loftus : A spreading activation theory of semantic processing. Psychological Review, 82:407-428 (1975)
3. C. Fellbaum (Ed.): WordNet: An Electronic Database. MIT Press, Cambridge, MA (1998)
4. W. N. Francis and H. Kucera: Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin (1982)
5. Evgeniy Gabrilovich and Shaul Markovitch: Computing semantic relatedness using wikipediabased explicit semantic analysis. In Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI 2007), pages 1606–1611, Hyderabad, January (2007)
6. Graeme Hirst and David St-Onge: Lexical chains as representations of context for the detection and correction of malapropisms. WordNet An electronic lexical database, pages 305–332, April (1998)

7. Thad Hughes and Daniel Ramage: Lexical semantic relatedness with random graph walks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing - Conference on Computational Natural Language Learning (EMNLP-CoNLL), pages 581–589, Prague, June (2007)
8. Dietmar Janetzko: Objectivity, Reliability, and Validity of Search Engine Count Estimates. International Journal of Internet Science, 3 (1), pages 7-33 (2008)
9. Mario Jarmasz and Stan Szpakowicz: Roget's thesaurus and semantic similarity. In Proceedings of Recent Advances in Natural Language Processing, pages 212-219, Borovets, September (2003)
10. Jay J Jiang and David W Conrath: Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, pages 19-33, Taipei, Taiwan, August (1997)
11. Claudia Leacock and Martin Chodorow: Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), pages 265–283 (1998)
12. Dekang Lin: An information-theoretic definition of similarity. In Proceedings of the 15th international conference on Machine Learning, volume 1, pages 296-304, Madison, July (1998)
13. R. Navigli: Word Sense Disambiguation: A survey. ACM Computing Surveys 41 (2009)
14. R. Navigli and S. Ponzetto: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier (2012)
15. Quillian, 1968 M. Ross Quillian : Semantic memory. In M. Minsky, editor, Semantic Information Processing. MIT Press, Cambridge, MA (1968)
16. Paul Rayson, Oliver Charles and Ian Auty: Can Google count? Estimating search engine result consistency. Proceedings of the seventh Web as Corpus Workshop, pages 23-30 (2012)
17. Philip Resnik: Using information content to evaluate semantic similarity in a taxonomy. In International Joint Conference for Artificial Intelligence, pages 448-453, Montreal, August (1995)
18. Reda Siblini and Leila Kosseim: Using a Weighted Semantic Network for Lexical Semantic Relatedness. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2013), September, Hissar, Bulgaria (2013)
19. Michael Strube and Simone Paolo Ponzetto: WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the National Conference on Artificial Intelligence, volume 21, page 1419, Boston, July (2006)
20. Egidio Terra and Charles LA Clarke: Frequency estimates for statistical word similarity measures. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, volume 21, pages 165-172, Edmonton, May (2003)
21. George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis: Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research, 37(1), pages 1–40 (2010)
22. Peter Turney: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Proceedings of the Twelfth European Conference on Machine Learning, pages 491-502, Freiburg Germany, September (2001)
23. Zhibiao Wu and Martha Palmer: Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138, New Mexico, June (1994)