Insights from Network Structure for Text Mining

Zornitsa Kozareva and Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{kozareva,hovy}@isi.edu

Abstract

Text mining and data harvesting algorithms have become popular in the computational linguistics community. They employ patterns that specify the kind of information to be harvested, and usually bootstrap either the pattern learning or the term harvesting process (or both) in a recursive cycle, using data learned in one step to generate more seeds for the next. They therefore treat the source text corpus as a network, in which words are the nodes and relations linking them are the edges. The results of computational network analysis, especially from the world wide web, are thus applicable. Surprisingly, these results have not yet been broadly introduced into the computational linguistics community. In this paper we show how various results apply to text mining, how they explain some previously observed phenomena, and how they can be helpful for computational linguistics applications.

1 Introduction

Text mining / harvesting algorithms have been applied in recent years for various uses, including learning of semantic constraints for verb participants (Lin and Pantel, 2002) related pairs in various relations, such as part-whole (Girju et al., 2003), cause (Pantel and Pennacchiotti, 2006), and other typical information extraction relations, large collections of entities (Soderland et al., 1999; Etzioni et al., 2005), features of objects (Pasca, 2004) and ontologies (Carlson et al., 2010). They generally start with one or more seed terms and employ patterns that specify the desired information as it relates to the seed(s). Several approaches have been developed specifically for learning patterns, including guided pattern collection with manual filtering (Riloff and Shepherd, 1997) automated surface-level pattern induction (Agichtein and Gravano, 2000; Ravichandran and Hovy, 2002) probabilistic methods for taxonomy relation learning (Snow et al., 2005) and kernel methods for relation learning (Zelenko et al., 2003). Generally, the harvesting procedure is recursive, in which data (terms or patterns) gathered in one step of a cycle are used as seeds in the following step, to gather more terms or patterns.

This method treats the source text as a graph or network, consisting of terms (words) as nodes and inter-term relations as edges. Each relation type induces a different network[1]. Text mining is a process of network traversal, and faces the standard problems of handling cycles, ranking search alternatives, estimating yield maxima, etc.

The computational properties of large networks and large network traversal have been studied intensively (Sabidussi, 1966; Freeman, 1979; Watts and Strogatz, 1998) and especially, over the past years, in the context of the world wide web (Page et al., 1999; Broder et al., 2000; Kleinberg and Lawrence, 2001; Li et al., 2005; Clauset et al., 2009). Surprisingly, except in (Talukdar and Pereira, 2010), this work has not yet been related to text mining research in the computational linguistics community.

The work is, however, relevant in at least two ways. It sometimes explains why text mining algo-

[1]These networks are generally far larger and more densely interconnected than the world wide web's network of pages and hyperlinks.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1616–1625,
Portland, Oregon, June 19-24, 2011.

rithms have the limitations and thresholds that are empirically found (or suspected), and it may suggest ways to improve text mining algorithms for some applications.

In Section 2, we review some related work. In Section 3 we describe the general harvesting procedure, and follow with an examination of the various statistical properties of implicit semantic networks in Section 4, using our implemented harvester to provide illustrative statistics. In Section 5 we discuss implications for computational linguistics research.

2 Related Work

The Natural Language Processing knowledge harvesting community has developed a good understanding of how to harvests various kinds of semantic information and use this information to improve the performance of tasks such as information extraction (Riloff, 1993), textual entailment (Zanzotto et al., 2006), question answering (Katz et al., 2003), and ontology creation (Suchanek et al., 2007), among others. Researchers have focused on the automated extraction of semantic lexicons (Hearst, 1992; Riloff and Shepherd, 1997; Girju et al., 2003; Pasca, 2004; Etzioni et al., 2005; Kozareva et al., 2008). While clustering approaches tend to extract general facts, pattern based approaches have shown to produce more constrained but accurate lists of semantic terms. To extract this information, (Lin and Pantel, 2002) showed the effect of using different sizes and genres of corpora such as news and Web documents. The latter has been shown to provide broader and more complete information.

Researchers outside computational linguistics have studied complex networks such as the World Wide Web, the Social Web, the network of scientific papers, among others. They have investigated the properties of these text-based networks with the objective of understanding their structure and applying this knowledge to determine node importance/centrality, connectivity, growth and decay of interest, etc. In particular, the ability to analyze networks, identify influential nodes, and discover hidden structures has led to important scientific and technological breakthroughs such as the discovery of communities of like-minded individuals (Newman and Girvan, 2004), the identification of influential people (Kempe et al., 2003), the ranking of scientists by their citation indexes (Radicchi et al., 2009), and the discovery of important scientific papers (Walker et al., 2006; Chen et al., 2007; Sayyadi and Getoor, 2009). Broder et al. (2000) demonstrated that the Web link structure has a "bow-tie" shape, while (2001) classified Web pages into *authorities* (pages with relevant information) and *hubs* (pages with useful references). These findings resulted in the development of the PageRank (Page et al., 1999) algorithm which analyzes the structure of the hyperlinks of Web documents to find pages with authoritative information. PageRank has revolutionized the whole Internet search society.

However, no-one has studied the properties of the text-based semantic networks induced by semantic relations between terms with the objective of understanding their structure and applying this knowledge to improve concept discovery. Most relevant to this theme is the work of Steyvers and Tenenbaum (Steyvers and Tenenbaum, 2004), who studied three manually built lexical networks (association norms, WordNet, and Roget's Thesaurus (Roget, 1911)) and proposed a model of the growth of the semantic structure over time. These networks are limited to the semantic relations among nouns.

In this paper we take a step further to explore the statistical properties of semantic networks relating *proper names*, *nouns*, *verbs*, and *adjectives*. Understanding the semantics of nouns, verbs, and adjectives has been of great interest to linguists and cognitive scientists such as (Gentner, 1981; Levin and Somers, 1993; Gasser and Smith, 1998). We implement a general harvesting procedure and show its results for these word types. A fundamental difference with the work of (Steyvers and Tenenbaum, 2004) is that we study very large semantic networks built 'naturally' by (millions of) users rather than 'artificially' by a small set of experts. The large networks capture the semantic intuitions and knowledge of the collective mass. It is conceivable that an analysis of this knowledge can begin to form the basis of a large-scale theory of semantic meaning and its interconnections, support observation of the process of lexical development and usage in humans, and even suggest explanations of how knowledge is organized in our brains, especially when performed for differ-

ent languages on the WWW.

3 Inducing Semantic Networks in the Web

Text mining algorithms such as those mentioned above raise certain questions, such as: *Why are some seed terms more powerful (provide a greater yield) than others?*, *How can one find high-yield terms?*, *How many steps does one need, typically, to learn all terms for a given relation?*, *Can one estimate the total eventual yield of a given relation?*, and so on. On the face of it, one would need to know the structure of the network a priori to be able to provide answers. But research has shown that some surprising regularities hold. For example, in the text mining community, (Kozareva and Hovy, 2010b) have shown that one can obtain a quite accurate estimate of the eventual yield of a pattern and seed after only five steps of harvesting. Why is this? They do not provide an answer, but research from the network community does.

To illustrate the properties of networks of the kind induced by semantic relations, and to show the applicability of network research to text harvesting, we implemented a harvesting algorithm and applied it to a representative set of relations and seeds in two languages.

Since the goal of this paper is not the development of a new text harvesting algorithm, we implemented a version of an existing one: the so-called DAP (doubly-anchored pattern) algorithm (Kozareva et al., 2008), because it (1) is easy to implement, (2) requires minimum input (one pattern and one seed example), (3) achieves very high precision compared to existing methods (Pasca, 2004; Etzioni et al., 2005; Pasca, 2007), (4) enriches existing semantic lexical repositories such as WordNet and Yago (Suchanek et al., 2007), (5) can be formulated to learn semantic lexicons and relations for *noun*, *verb* and *verb+preposition* syntactic constructions; (6) functions equally well in different languages. Next we describe the knowledge harvesting procedure and the construction of the text-mined semantic networks.

3.1 Harvesting to Induce Semantic Networks

For a given semantic class of interest say *singers*, the algorithm starts with a *seed* example of the *class*, say *Madonna*. The *seed* term is inserted in the lexico-syntactic pattern "*class* such as *seed* and *", which learns on the position of the $*$ new terms of type *class*. The newly learned terms are then individually placed into the position of the *seed* in the pattern, and the bootstrapping process is repeated until no new terms are found. The output of the algorithm is a set of terms for the semantic class. The algorithm is implemented as a breadth-first search and its mechanism is described as follows:

1. Given:
   a language L={English, Spanish}
   a pattern P_i={such as, including, verb prep, noun}
   a seed term $seed$ for P_i
2. Build a query for P_i using template T_i 'class such as $seed$ and *', 'class including $seed$ and *', '* and $seed$ verb prep', '* and $seed$ noun', '$seed$ and * noun'
3. Submit T_i to Yahoo! or other search engine
4. Extract terms occupying the * position
5. Feed terms from 4. into 2.
6. Repeat steps 2–5. until no new terms are found

The output of the knowledge harvesting algorithm is a network of semantic terms interconnected by the semantic relation captured in the pattern. We can represent the traversed (implicit) network as a directed graph $G(V, E)$ with nodes $V(|V| = n)$ and edges $E(|E| = m)$. A node u in the network corresponds to a term discovered during bootstrapping. An edge $(u, v) \in E$ represents an existing link between two terms. The direction of the edge indicates that the term v was generated by the term u. For example, given the sentence (where the pattern is in italics and the extracted term is underlined) "He loves *singers such as Madonna and* Michael Jackson", two nodes *Madonna* and *Michael Jackson* with an edge e=(*Madonna*, *Michael Jackson*) would be created in the graph G. Figure 1 shows a small example of the singer network. The starting seed term *Madonna* is shown in red color and the harvested terms are in blue.

3.2 Data

We harvested data from the Web for a representative selection of semantic classes and relations, of

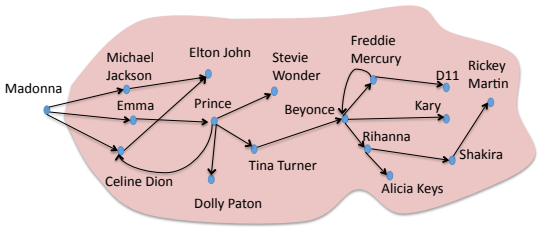

Figure 1: Harvesting Procedure.

the type used in (Etzioni et al., 2005; Pasca, 2007; Kozareva and Hovy, 2010a):

- semantic classes that can be learned using different seeds (e.g., "singers such as **Madonna** and *" and "singers such as **Placido Domingo** and *");
- semantic classes that are expressed through different lexico-syntactic patterns (e.g., "weapons **such as** bombs and *" and "weapons **including** bombs and *");
- verbs and adjectives characterizing the semantic class (e.g., "**expensive** and * car", "dogs **run** and *");
- semantic relations with more complex lexico-syntactic structure (e.g., "* and Easyjet **fly to**", "* and Sam **live in**");
- semantic classes that are obtained in different languages, such as English and Spanish (e.g., "**singers** such as Madonna and *" and "**cantantes** como Madonna y *");

While most of these variations have been explored in individual papers, we have found no paper that covers them all, and none whatsoever that uses verbs and adjectives as seeds.

Using the above procedure to generate the data, each pattern was submitted as a query to Yahoo!Boss. For each query the top 1000 text snippets were retrieved. The algorithm ran until exhaustion. In total, we collected 10GB of data which was part-of-speech tagged with Treetagger (Schmid, 1994) and used for the semantic term extraction. Table 1 summarizes the number of nodes and edges learned for each semantic network using pattern P_i and the initial seed shown in italics.

Lexico-Syntactic Pattern	Nodes	Edges
P_1="**singers** such as *Madonna* and *"	1115	1942
P_2="**singers** such as *Placido Domingo* and *"	815	1114
P_3="**emotions** including *anger* and *"	113	250
P_4="**emotions** such as *anger* and *"	748	2547
P_5="**diseases** such as *malaria* and *"	3168	6752
P_6="**drugs** such as *ibuprofen* and *"	2513	9428
P_7="*expensive* and * **cars**"	4734	22089
P_8="* and *tasty* **fruits**"	1980	7874
P_9="**whales** *swim* and *"	869	2163
P_{10}="**dogs** *chase* and *"	4252	20212
P_{11}="**Britney Spears** *dances* and *"	354	540
P_{12}="**John** *reads* and *"	3894	18545
P_{13}="* and *Easyjet* **fly to**"	3290	6480
P_{14}="* and *Charlie* **work for**"	2125	3494
P_{15}="* and *Sam* **live in**"	6745	24348
P_{16}="**cantantes** como *Madonna* y *"	240	318
P_{17}="**gente** como *Jorge* y *"	572	701

Table 1: *Size of the Semantic Networks.*

4 Statistical Properties of Text-Mined Semantic Networks

In this section we apply a range of relevant measures from the network analysis community to the networks described above.

4.1 Centrality

The first statistical property we explore is centrality. It measures the degree to which the network structure determines the importance of a node in the network (Sabidussi, 1966; Freeman, 1979).

We explore the effect of two centrality measures: $indegree$ and $outdegree$. The $indegree$ of a node u denoted as $indegree(u)$=$\sum(v,u)$ considers the sum of all incoming edges to u and captures the ability of a semantic term to be discovered by other semantic terms. The $outdegree$ of a node u denoted as $outdegree(u)$=$\sum(u,v)$ considers the number of outgoing edges of the node u and measures the ability of a semantic term to discover new terms. Intuitively, the more central the node u is, the more confident we are that it is a correct term.

Since harvesting algorithms are notorious for extracting erroneous information, we use the two centrality measures to rerank the harvested elements. Table 2 shows the accuracy[2] of the singer semantic terms at different ranks using the in and out degree measures. Consistently, $outdegree$ outperforms $indegree$ and reaches higher accuracy. This

[2]Accuracy is calculated as the number of correct terms at rank R divided by the total number of terms at rank R.

shows that for the text-mined semantic networks, the ability of a term to discover new terms is more important than the ability to be discovered.

@rank	in-degree	out-degree
10	.92	1.0
25	.91	1.0
50	.90	.97
75	.90	.96
100	.89	.96
150	.88	.95

Table 2: *Accuracy of the Singer Terms.*

This poses the question "*What are the terms with high and low outdegree?*". Table 3 shows the top and bottom 10 terms of the semantic class.

Semantic Class	top 10 outDegree	bottom 10 outDegree
Singers	Frank Sinatra	Alanis Morisette
	Ella Fitzgerald	Christine Agulera
	Billie Holiday	Buffy Sainte-Marie
	Britney Spears	Cece Winans
	Aretha Franklin	Wolfman Jack
	Michael Jackson	Billie Celebration
	Celine Dion	Alejandro Sanz
	Beyonce	France Gall
	Bessie Smith	Peter
	Joni Mitchell	Sarah

Table 3: *Singer Term Ranking with Centrality Measures.*

The nodes with high outdegree correspond to famous or contemporary singers. The lower-ranked nodes are mostly spelling errors such as *Alanis Morisette* and *Christine Agulera*, less known singers such as *Buffy Sainte-Marie* and *Cece Winans*, non-American singers such as *Alejandro Sanz* and *France Gall*, extractions due to part-of-speech tagging errors such as *Billie Celebration*, and general terms such as *Peter* and *Sarah*. Potentially, knowing which terms have a high *outdegree* allows one to rerank candidate seeds for more effective harvesting.

4.2 Power-law Degree Distribution

We next study the degree distributions of the networks. Similarly to the Web (Broder et al., 2000) and social networks like Orkut and Flickr, the text-mined semantic networks also exhibit a power-law distribution. This means that while a few terms have a significantly high degree, the majority of the semantic terms have small degree. Figure 2 shows the $indegree$ and $outdegree$ distributions for different semantic classes, lexico-syntactic patterns, and languages (English and Spanish). For each semantic network, we plot the best-fitting power-law function (Clauset et al., 2009) which fits well all degree distributions. Table 4 shows the power-law exponent values for all text-mined semantic networks.

Patt.	γ_{in}	γ_{out}	Patt.	γ_{in}	γ_{out}
P_1	2.37	1.27	P_{10}	1.65	1.12
P_2	2.25	1.21	P_{11}	2.42	1.41
P_3	2.20	1.76	P_{12}	1.60	1.13
P_4	2.28	1.18	P_{13}	2.26	1.20
P_5	2.49	1.18	P_{14}	2.43	1.25
P_6	2.42	1.30	P_{15}	2.51	1.43
P_7	1.95	1.20	P_{16}	2.74	1.31
P_8	1.94	1.07	P_{17}	2.90	1.20
P_9	1.96	1.30			

Table 4: *Power-Law Exponents of Semantic Networks.*

It is interesting to note that the $indegree$ power-law exponents for all semantic networks fall within the same range ($\gamma_{in} \approx 2.4$), and similarly for the $outdegree$ exponents ($\gamma_{out} \approx 1.3$). However, the values of the $indegree$ and $outdegree$ exponents differ from each other. This observation is consistent with Web degree distributions (Broder et al., 2000). The difference in the distributions can be explained by the link asymmetry of semantic terms: A discovering B does not necessarily mean that B will discover A. In the text-mined semantic networks, this asymmetry is caused by patterns of language use, such as the fact that people use first adjectives of the size and then of the color (e.g., big red car), or prefer to place male before female proper names. Harvesting patterns should take into account this tendency.

4.3 Sparsity

Another relevant property of the semantic networks concerns sparsity. Following Preiss (Preiss, 1999), a graph is sparse if $|E| = O(|V|^k)$ and $1 < k < 2$, where $|E|$ is the number of edges and $|V|$ is the number of nodes, otherwise the graph is dense. For the studied text-semantic networks, k is ≈ 1.08. Sparsity can be also captured through the density of the semantic network which is computed as $\frac{|E|}{V(V-1)}$. All networks have low density which suggests that the networks exhibit a sparse connectivity pattern. On average a node (semantic term) is connected to a very small percentage of other nodes. Similar behavior was reported for the WordNet and Roget's semantic networks (Steyvers and Tenenbaum, 2004).

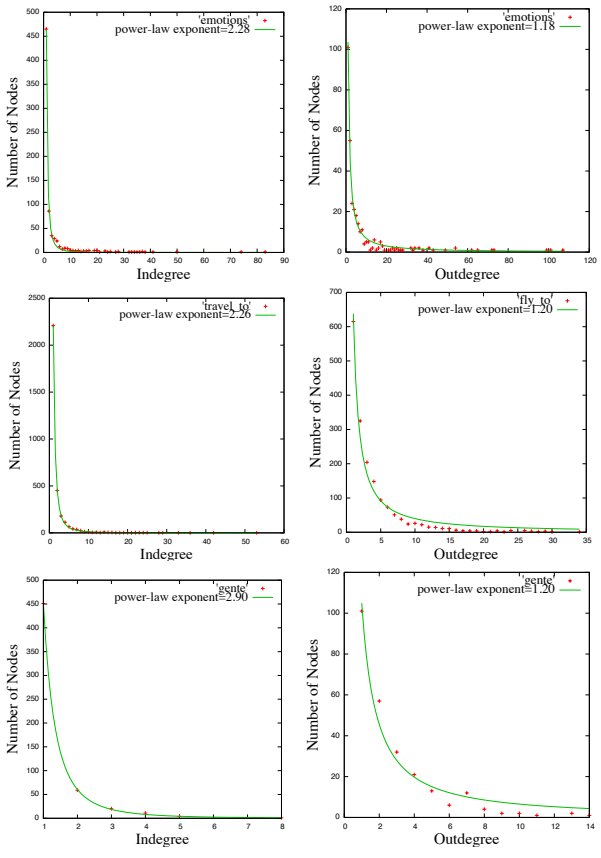

Figure 2: Degree Distributions of Semantic Networks.

4.4 Connectedness

For every network, we computed the strongly connected component (SCC) such that for all nodes (semantic terms) in the SCC, there is a path from any node to another node in the SCC considering the direction of the edges between the nodes. For each network, we found that there is only one SCC. The size of the component is shown in Table 5. Unlike WordNet and Roget's semantic networks where the SCC consists 96% of all semantic terms, in the text-mined semantic networks only 12 to 55% of the terms are in the SCC. This shows that not all nodes can reach (discover) every other node in the network. This also explains the findings of (Kozareva et al., 2008; Vyas et al., 2009) why starting with a good seed is important.

4.5 Path Lengths and Diameter

Next, we describe the properties of the shortest paths between the semantic terms in the SCC. The distance between two nodes in the SCC is measured as the length of the shortest path connecting the terms. The direction of the edges between the terms is taken into consideration. The average distance is the average value of the shortest path lengths over all pairs of nodes in the SCC. The diameter of the SCC is calculated as the maximum distance over all pairs of nodes (u, v), such that a node v is reachable from node u. Table 5 shows the average distance and the diameter of the semantic networks.

Patt.	#nodes in SCC	SCC Average Distance	SCC Diameter
P_1	364 (.33)	5.27	16
P_2	285 (.35)	4.65	13
P_3	48 (.43)	2.85	6
P_4	274 (.37)	2.94	7
P_5	1249 (.38)	5.99	17
P_6	1471 (.29)	4.82	15
P_7	2255 (.46)	3.51	11
P_8	1012 (.50)	3.87	11
P_9	289 (.33)	4.93	13
P_{10}	2342 (.55)	4.50	12
P_{11}	87 (.24)	5.00	11
P_{12}	1967 (.51)	3.20	13
P_{13}	1249 (.38)	4.75	13
P_{14}	608 (.29)	7.07	23
P_{15}	1752 (.26)	5.32	15
P_{16}	56 (.23)	4.79	12
P_{17}	69 (.12)	5.01	13

Table 5: *SCC, SCC Average Distance and SCC Diameter of the Semantic Networks.*

The diameter shows the maximum number of steps necessary to reach from any node to any other, while the average distance shows the number of steps necessary on average. Overall, all networks have very short average path lengths and small diameters that are consistent with Watt's finding for small-world networks. Therefore, the yield of harvesting seeds can be predicted within five steps explaining (Kozareva and Hovy, 2010b; Vyas et al., 2009).

We also compute for any randomly selected node in the semantic network on average how many hops (steps) are necessary to reach from one node to another. Figure 3 shows the obtained results for some of the studied semantic networks.

4.6 Clustering

The clustering coefficient (C) is another measure to study the connectivity structure of the networks (Watts and Strogatz, 1998). This measure captures the probability that the two neighbors of a randomly selected node will be neighbors. The clustering coefficient of a node u is calculated as $C_u = \frac{|e_{ij}|}{k_u(k_u-1)}$

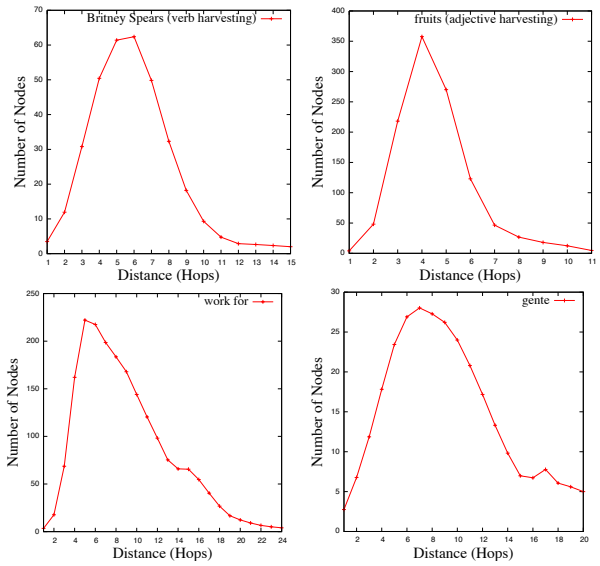

Figure 3: Hop Plot of the Semantic Networks.

Patt.	C	ClosedTriads	OpenTriads
P_1	.01	14096 (.97)	388 (.03)
P_2	.01	6487 (.97)	213 (.03)
P_3	.30	1898 (.94)	129 (.06)
P_4	.33	60734 (.94)	3944 (.06)
P_5	.10	79986 (.97)	2321 (.03)
P_6	.11	78716 (.97)	2336 (.03)
P_7	.17	910568 (.95)	43412 (.05)
P_8	.19	21138 (.95)	10728 (.05)
P_9	.20	27830 (.95)	1354 (.05)
P_{10}	.15	712227 (.96)	62101(.04)
P_{11}	.09	3407 (.98)	63 (.02)
P_{12}	.15	734724 (.96)	32517 (.04)
P_{13}	.06	66162 (.99)	858 (.01)
P_{14}	.05	28216 (.99)	408 (.01)
P_{15}	.09	1336679 (.97)	47110 (.03)
P_{16}	.09	1525 (.98)	37 (.02)
P_{17}	.05	2222 (.99)	21 (.01)

Table 6: *Clustering Coefficient of the Semantic Networks.*

: $v_i, v_j \in N_u, e_{ij} \in E$, where k_u is the total degree of the node u and N_u is the neighborhood of u. The clustering coefficient C for the whole semantic network is the average clustering coefficient of all its nodes, C=$\frac{1}{n}\sum C_i$. The value of the clustering coefficient ranges between $[0, 1]$, where 0 indicates that the nodes do not have neighbors which are themselves connected, while 1 indicates that all nodes are connected. Table 6 shows the clustering coefficient for all text-mined semantic networks together with the number of closed and open triads[3]. The analysis suggests the presence of a strong local cluster, however there are few possibilities to form overlapping neighborhoods of nodes. The clustering coefficient of WordNet (Steyvers and Tenenbaum, 2004) is similar to those of the text-mined networks.

4.7 Joint Degree Distribution

In social networks, understanding the preferential attachment of nodes is important to identify the speed with which epidemics or gossips spread. Similarly, we are interested in understanding how the nodes of the semantic networks connect to each other. For this purpose, we examine the Joint Degree Distribution (JDD) (Li et al., 2005; Newman, 2003). JDD is approximated by the degree correlation function k_{nn} which maps the $outdegree$ and the average $indegree$ of all nodes connected to a node with that $outdegree$. High values of k_{nn} indicate that high-degree nodes tend to connect to other high-degree nodes (forming a "core" in the network), while lower values of k_{nn} suggest that the high-degree nodes tend to connect to low-degree ones. Figure 4 shows the k_{nn} for the *singer*, *whale*, *live in*, *cars*, *cantantes*, and *gente* networks. The figure plots the outdegree and the average indegree of the semantic terms in the networks on a log-log scale. We can see that for all networks the high-degree nodes tend to connect to other high-degree ones. This explains why text mining algorithms should focus their effort on high-degree nodes.

4.8 Assortivity

The property of the nodes to connect to other nodes with similar degrees can be captured through the assortivity coefficient r (Newman, 2003). The range of r is $[-1, 1]$. A positive assortivity coefficient means that the nodes tend to connect to nodes of similar degree, while negative coefficient means that nodes are likely to connect to nodes with degree very different from their own. We find that the assortivitiy coefficient of our semantic networks is positive, ranging from 0.07 to 0.20. In this respect, the semantic networks differ from the Web, which has a negative assortivity (Newman, 2003). This implies a difference in text mining and web search traversal strategies: since starting from a highly-connected seed term will tend to lead to other highly-connected terms, text mining algorithms should prefer depth-first traversal, while web search algorithms starting

[3] A triad is three nodes that are connected by either two (open triad) or three (closed triad) directed ties.

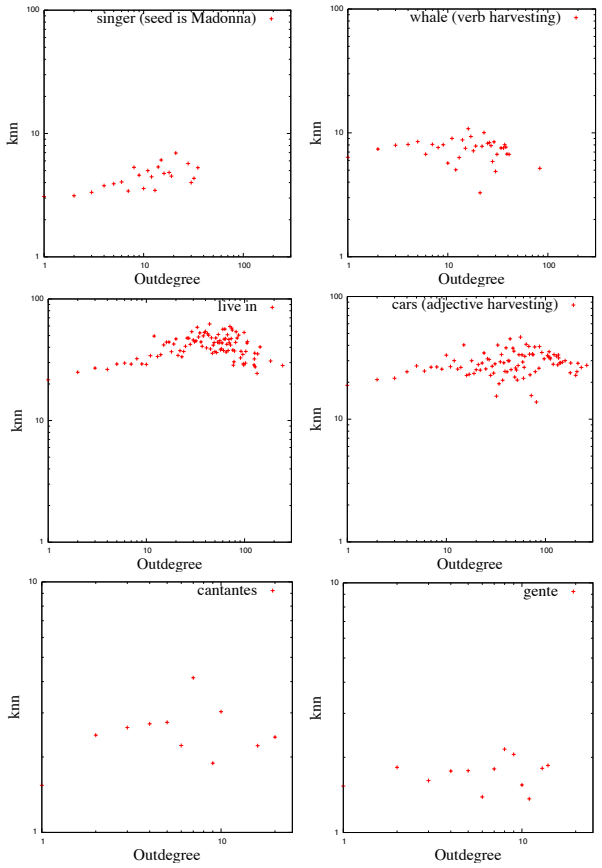

Figure 4: Joint Degree Distribution of the Semantic Networks.

from a highly-connected seed page should prefer a breadth-first strategy.

5 Discussion

The above studies show that many of the properties discovered of the network formed by the web hold also for the networks induced by semantic relations in text mining applications, for various semantic classes, semantic relations, and languages. We can therefore apply some of the research from network analysis to text mining.

The small-world phenomenon, for example, holds that any node is connected to any other node in at most six steps. Since as shown in Section 4.5 the semantic networks also exhibit this phenomenon, we can explain the observation of (Kozareva and Hovy, 2010b) that one can quite accurately predict the relative 'goodness' of a seed term (its eventual total yield and the number of steps required to obtain that) within five harvesting steps. We have shown that due to the strongly connected components in text mining networks, not all elements within the harvested graph can discover each other. This implies that harvesting algorithms have to be started with several seeds to obtain adequate Recall (Vyas et al., 2009). We have shown that centrality measures can be used successfully to rank harvested terms to guide the network traversal, and to validate the correctness of the harvested terms.

In the future, the knowledge and observations made in this study can be used to model the lexical usage of people over time and to develop new semantic search technology.

6 Conclusion

In this paper we describe the implicit 'hidden' semantic network graph structure induced over the text of the web and other sources by the semantic relations people use in sentences. We describe how term harvesting patterns whose seed terms are harvested and then applied recursively can be used to discover these semantic term networks. Although these networks differ considerably from the web in relation density, type, and network size, we show, somewhat surprisingly, that the same power-law, small-world effect, transitivity, and most other characteristics that apply to the web's hyperlinked network structure hold also for the implicit semantic term graphs—certainly for the semantic relations and languages we have studied, and most probably for almost all semantic relations and human languages.

This rather interesting observation leads us to surmise that the hyperlinks people create in the web are of essentially the same type as the semantic relations people use in normal sentences, and that they form an extension of normal language that was not needed before because people did not have the ability within the span of a single sentence to 'embed' structures larger than a clause—certainly not a whole other page's worth of information. The principal exception is the academic citation reference (lexicalized as "see"), which is not used in modern webpages. Rather, the 'lexicalization' now used is a formatting convention: the hyperlink is colored and often underlined, facilities offered by computer screens but not available to speech or easy in traditional typesetting.

Acknowledgments

We acknowledge the support of DARPA contract number FA8750-09-C-3705 and NSF grant IIS-0429360. We would like to thank Sujith Ravi for his useful comments and suggestions.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. pages 85–94.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the web. *Comput. Netw.*, 33(1-6):309–320.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. pages 101–110.

Peng Chen, Huafeng Xie, Sergei Maslov, and Sid Redner. 2007. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, January.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134, June.

Linton Freeman. 1979. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Michael Gasser and Linda B. Smith. 1998. Learning nouns and adjectives: A connectionist account. In *Language and Cognitive Processes*, pages 269–306.

Demdre Gentner. 1981. Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, pages 161–178.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 1–8.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545.

Boris Katz, Jimmy Lin, Daniel Loreto, Wesley Hildebrandt, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton, and Federico Mora. 2003. Integrating web-based and corpus-based techniques for question answering. In *Proceedings of the twelfth text retrieval conference (TREC)*, pages 426–435.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146.

Jon Kleinberg and Steve Lawrence. 2001. The structure of the web. *Science*, 29:1849–1850.

Zornitsa Kozareva and Eduard Hovy. 2010a. Learning arguments and supertypes of semantic relations using recursive patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010*, pages 1482–1491, July.

Zornitsa Kozareva and Eduard Hovy. 2010b. Not all seeds are equal: Measuring the quality of text mining seeds. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 618–626.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics ACL-08: HLT*, pages 1048–1056.

Beth Levin and Harold Somers. 1993. English verb classes and alternations: A preliminary investigation.

Lun Li, David Alderson, Reiko Tanaka, John C. Doyle, and Walter Willinger. 2005. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version). *Internet Mathematica*, 2(4):431–523.

Dekang Lin and Patrick Pantel. 2002. Concept discovery from text. In *Proc. of the 19th international conference on Computational linguistics*, pages 1–7.

Mark E. Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review*, 69(2).

Mark Newman. 2003. Mixing patterns in networks. *Physical Review E*, 67.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. pages 113–120.

Marius Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 137–145.

Marius Pasca. 2007. Weakly-supervised discovery of named entities using web search queries. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007*, pages 683–690.

Bruno R. Preiss. 1999. *Data structures and algorithms with object-oriented design patterns in C++*.

Filippo Radicchi, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. 2009. Diffusion of scientific credits and the ranking of scientists. In *Phys. Rev. E 80, 056103*.

Deepack Ravichandran and Eduard H. Hovy. 2002. Learning surface text patterns for a question answering system. pages 41–47.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Empirical Methods for Natural Language Processing*, pages 117–124.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. pages 811–816.

Peter Mark Roget. 1911. *Roget's thesaurus of English Words and Phrases*. New York Thomas Y. Crowell company.

Gert Sabidussi. 1966. The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Hassan Sayyadi and Lise Getoor. 2009. Future rank: Ranking scientific articles by predicting their future pagerank. In *2009 SIAM International Conference on Data Mining (SDM09)*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. pages 1297–1304.

Stephen Soderland, Claire Cardie, and Raymond Mooney. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning, 34(1-3)*, pages 233–272.

Mark Steyvers and Joshua B. Tenenbaum. 2004. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29:41–78.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Partha Pratim Talukdar and Fernando Pereira. 2010. Graph-based weakly-supervised methods for information extraction and integration. pages 1473–1481.

Vishnu Vyas, Patrick Pantel, and Eric Crestan. 2009. Helping editors choose better seed sets for entity set expansion. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM*, pages 225–234.

Dylan Walker, Huafeng Xie, Koon-Kiu Yan, and Sergei Maslov. 2006. Ranking scientific publications using a simple model of network traffic. December.

Duncan Watts and Steven Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 849–856.

Dmitry Zelenko, Chinatsu Aone, Anthony Richardella, Jaz K, Thomas Hofmann, Tomaso Poggio, and John Shawe-taylor. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research 3*.