Chinese Science Bulletin

© 2009 SCIENCE IN CHINA PRESS

Springer

Statistical properties of Chinese semantic networks

LIU HaiTao

Institute of Applied Linguistics, Communication University of China, Beijing 100024, China

Almost all language networks in word and syntactic levels are small-world and scale-free. This raises the questions of whether a language network in deeper semantic or cognitive level also has the similar properties. To answer the question, we built up a Chinese semantic network based on a treebank with semantic role (argument structure) annotation and investigated its global statistical properties. The results show that although semantic network is also small-world and scale-free, it is different from syntactic network in hierarchical structure and K-Nearest-Neighbor correlation.

semantic network, semantic role, complex network, Chinese, small-world, scale-free

Language networks are small-world and scale-free, although they are built based on different principles[1]. Similar global statistical properties, which are shown by language networks, are independent of linguistic structure and typology[1–5]. If the global properties of language network could not reflect the differences of these structures, how could we consider that these statistical properties are indicators of a language network? Do linguistic structures really influence the statistical properties of a language network? More concretely, does syntactic network have the same properties with semantic or conceptual one? To answer the questions, it seems necessary to investigate the language network based on different linguistic principles or levels. Syntactic networks have been explored in several languages[2,4,5], but the statistical properties of (dynamic) semantic (argument structure) network based on real text have not been reported yet.

The study reported in this paper will explore these questions. To investigate statistical properties of semantic network, we built a corpus with semantic role (argument structure) annotation. The final corpus includes 34435 word tokens. Based on the corpus, we built a Chinese semantic network with 5903 nodes.

Considering the close relation between syntactic and semantic structures in a language, it is interesting to observe their differences and similarities from a view of complex network. In a semantic (language) network, a node represents an auto-semantic word, and the edge refers to the semantic relation between two words. Semantic network is an intermediate between syntactic and conceptual network. Therefore semantic networks, in particular, dynamic semantic networks (i.e. based on real language usage or text), are useful to explore the following three questions: the organization of human semantic (or conceptual) knowledge, human performance in semantic processing and the processes of semantic retrieval and search.

1 Methods

The semantic analysis (annotation) is structural and dependency based[6]. It captures the so-called deep, semantic structure of the sentence. While syntactic analysis has to link all words in a sentence into an integrated whole, semantic analysis only concerns the relations between auto-semantic (content) words. Semantic structure (annotation) in this study is similar to the structural analysis in Tesnière’s theory[7], tectogrammatical layer in Prague dependency treebank[8], DSyntS (deep syntactic structure) in meaning-text theory[9] and word bank in database semantics[10].

Figure 1 provides the syntactic and semantic analysis of the sentence ‘约翰在桌子上放了三本书’ (John puts three books on the table).

Received November 21, 2008; accepted March 23, 2009
doi: 10.1007/s11434-009-0467-x
email: lhtcuc@gmail.com
Supported by the National Social Science Foundation of China (Grant No. 09BYY024) and “211” Key Projects of Communication University of China

Citation: Liu H T. Statistical properties of Chinese semantic networks. Chinese Sci Bull, 2009, 54: 2781—2785, doi: 10.1007/s11434-009-0467-x

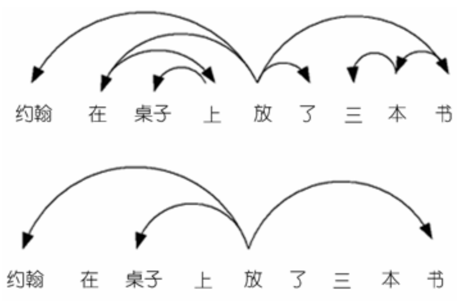

Figure 1 Syntactic and semantic dependency analysis of '约翰在桌子上放了三本书'. Upper is syntactic, below semantic.

Functional words do not play any role in semantic analysis. Therefore semantic network, which is based on semantic analysis, does not include functional words that are often the most important nodes or hubs in syntactic network. If semantic network excludes functional words, it should display different properties from syntactic network. On the other hand, considering the similarities between semantic analysis and conceptual graph[11], we can find some statistical properties of conceptual (cognitive) network by investigating semantic network.

Figure 2 shows such difference in an example network, which consists of three Chinese sentences: 约翰在桌子上放了本书 (John puts the book on the table), 那学生读过一本有趣的书 (The student read an interesting book), 那本书的封面旧了 (The cover of the book is old).

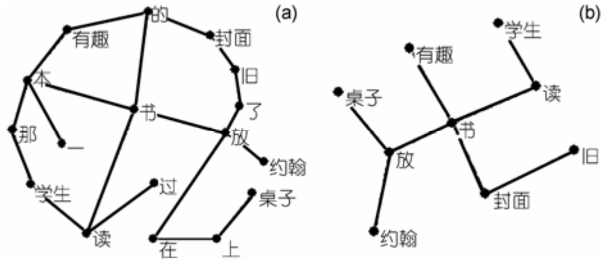

Figure 2 Syntactic and semantic networks of three Chinese sentences. (a) Syntactic; (b) semantic.

Figure 2(a) clearly shows the importance of functional words as linking nodes in syntactic network and Figure 2(b) is forming a simpler network. Figure 2 also shows that we cannot get semantic network directly from syntactic network by removing the nodes of functional words. For instance, if functional words (的，本，在，了) are removed from Figure 2(a), the network will be separated into four subparts. Therefore dynamic semantic network should be built based on semantic analysis of a text.

The network analysis software Pajek was used to extract the centers of the example networks (net→vector→centers) as in Figure 3.

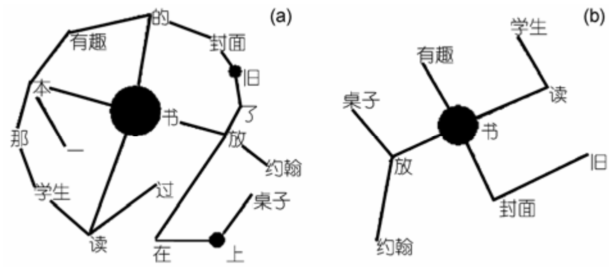

Figure 3 Centers in syntactic and semantic networks of three Chinese sentences. (a) Syntactic; (b) semantic.

A working corpus for this study is built from the news (xinwen lianbo) of China Central Television. We manually annotated the working corpus by the scheme on semantic annotation[12]. Final corpus with semantic annotation includes 34435 word tokens in 1486 Chinese sentences. Based on the method proposed in ref. [4], we converted the corpus into an undirected Chinese semantic network.

Figures 2 and 3 clearly show that syntactic and semantic analysis may influence the properties of a language network. However, do they make evident differences from the views of a complex network? Answers to this question are discussed in the section as follows.

2 Results

The average path length, the clustering coefficients and the degree distribution of a network are among the network indicators most frequently investigated for describing essential properties of the complexity of a network[13]. In this section, these three indicators of semantic network are calculated and discussed.

In a semantic (language) network, a node represents an auto-semantic word, and the edge refers to the semantic relation between two words. Average path length $\langle d \rangle$ is the average shortest distance between any pair of nodes in a network. Semantic network's $\langle d \rangle$ is 3.952. The maximum shortest path in a network is defined as diameter D of the network. Here it is 17.

The distribution of the shortest path lengths (Figure 4) is drawn based on the histogram of the length of the shortest paths between pairs of nodes of a network.

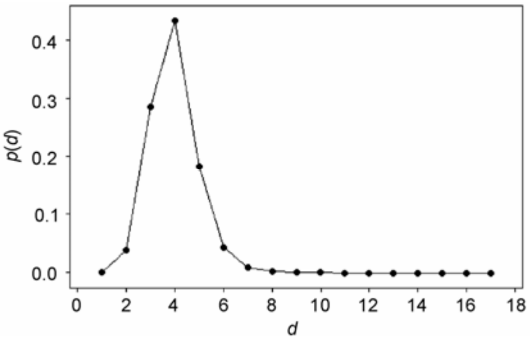

Figure 4 Shortest path distribution of semantic network.

Shortest path distribution of semantic network has a longer tail than that of syntactic network. Lack of functional words seems to make semantic network have greater ⟨d⟩ and D than syntactic networks[4].

In a language network, the number of links of a given word type is called its degree k. ⟨k⟩ is the average degree of a network, which reflects the combining capacity of a word with other words. The ⟨k⟩ of semantic network is 7.546, which is similar to syntactic network. Degree distributions are defined as the frequency $P(k)$ of having a word type with k links. Cumulative degree distribution of the semantic network is shown in Figure 5.

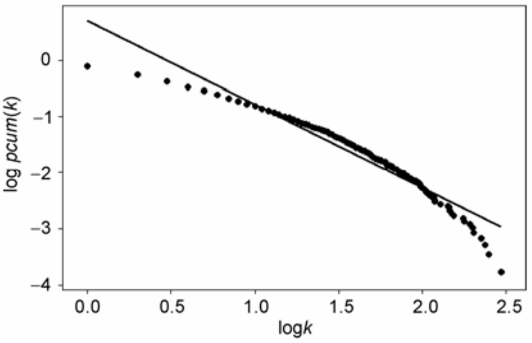

Figure 5 Cumulative degree distributions of semantic network. The cumulative degree distribution was fitted by a power law with slope of –1.493, which corresponds to the exponent γ = 2.493.

Clustering coefficient C measures the average probability that two neighbors of the same node (word) are also connected between them. Let k_i denote the degree of node i, and E_i denote the number of edges among the nodes in the nearest neighborhood of node i. Then the clustering coefficient C_i of the node i is $2E_i/k_i(k_i-1)$. The clustering coefficient of the network is given by the average of C_i over all the nodes in the network. In this case, it is 0.0794, which is smaller than syntactic network.

These essential complexity properties of semantic and syntactic networks are summarized in Table 1. Data of syntactic networks are from ref. [4].

Table 1 Main properties of semantic and syntactic networks[a)]

Network	N	⟨k⟩	C	⟨d⟩	D	γ	C_{rand}	⟨d_{rand}⟩
Semantic	5903	7.46	0.079	3.952	17	2.49	0.0011	4.55
Syntactic 1	4017	6.48	0.128	3.372	10	2.40	0.0014	4.66
Syntactic 2	2637	8.91	0.260	2.996	10	2.18	0.0036	3.83

a) N, number of nodes; <k>, average degree; C, clustering coefficient; <d>, average path length; D, diameter; γ, exponent of power law; C_{rand}, clustering coefficient of random graph; <d_{rand}>, average path length of random graph. Network “Syntactic 1” has the same text genre with “Semantic”, and “Syntactic 2” is based on conversational text.

If a network has a high clustering coefficient C (⟨d⟩~⟨d_{rand}⟩) and a very short path length ⟨d⟩ (C>>C_{rand}), it is a small-world (SW) network[14]. Following this criterion, the semantic network seems small-world, although it has smaller C than syntactic network. If the degree distribution of a network is following a power law,

$$P(k) \sim k^{-\gamma}, \tag{1}$$

and the constant γ is between 2 and 3, the network is a scale-free network[15]. Therefore as shown in Figure 5, the semantic network is scale-free.

Table 1 shows the difference between Semantic and Syntactic 1, which is smaller than that between Syntactic 1 and Syntactic 2. This is an interesting finding, although we do not clearly know which factors make that. Syntactic 1 and Semantic networks are built based on the corpora with the same genre (news), while Syntactic 2 uses conversational genre.

The property on hierarchical organization of a network can be measured in the correlation between the clustering coefficient and the degree of the nodes of a network. The correlation is expressed through the function $C(k)$, which represents the average clustering coefficient of all nodes with degree k[16]. Formally, it is defined as the probability that two nodes, neighbors of a node of degree k, are linked to each other. Thus it can be formulated as a function of the three nodes correlations:

$$\overline{C}(k) = \frac{1}{N_k}\sum_i C_i \delta_{k_i,k}, \tag{2}$$

where N_k is the total number of nodes with degree k; the sum runs over all possible nodes and $\delta_{ki,k}$ is the Kronecker delta, which has values $\delta_{i,j}$ = 1, if i=j and $\delta_{i,j}$ =0 if $i \neq j$.

In many real networks, $C(k)$ exhibits a highly significant behavior with a power-law decay as a function of k that signals a hierarchy in which most low degree nodes belong to well interconnected communities, and hubs connect many nodes that are not directly connected[17].

Figure 6 shows that the distributions of $C(k)$ are very skewed, which are not power laws as in other networks. The neighbors of a node with degree 1 mostly do not link to each other. In other words, a semantic network tends to create a longer path length between two nodes and a greater diameter than syntactic networks in ref. [2]. That makes semantic network a poorer hierarchy.

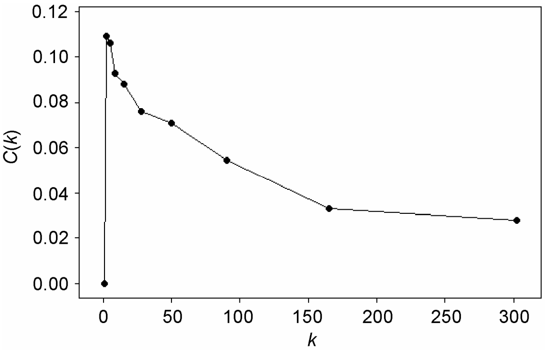

Figure 6 Clustering coefficient $C(k)$ vs. degree k for the semantic network.

Another indicator which can characterize the real-world networks, is K-Nearest-Neighbor (average neighbors degree) k_{NN} that measures the correlation between the degree of a node and that of its neighbors. A network is assortative mixing or assortativity if large (small) degree nodes tend to be linked with large (small) degree nodes. A network is disassortative mixing or disassortativity if large (small) degree nodes tend to be linked with small (large) degree nodes. Social networks are typical representatives of assortative mixing networks. Biological and technological networks are examples of disassortative mixing networks[18].

Refs. [17,19] have argued that the correct mathematical way to quantify such a measure is the conditional probability $P(k'|k)$ of having a node with degree k' at one side of the edge given that on the other side of the edge the degree is k. The degree correlation function can be investigated by the average degree of the nearest neighbors of nodes of degree k as follows:

$$\bar{k}_{NN}(k) = \sum_{k'} k' P(k'|k). \tag{3}$$

We used a less strict but more intuitive and simple approach proposed by ref. [20], which defined undirected K-Nearest-Neighbor as follows: a node is selected and the average degree of its all neighbors is calculated. By repeating the procedure for all nodes of the network, one derives a pair (k_{NN}, k) for each node, where k is the degree of the node. By averaging over nodes with equal degree k, one derives the function $k_{NN}(k)$, which allows to study the correlation. If $k_{NN}(k)$ grows with k, the network is assortative; if $k_{NN}(k)$ decreases with k, the network is disassortative. A flat curve would indicate the absence of correlation.

Figure 7 shows that there is a weaker correlation between k_{NN} and k in a semantic network than in a syntactic network[2]. The disassortative property of a syntactic network can reflect the relation between content and functional words. As a result, the absence of functional words makes a flatter curve in semantic network. Ref. [2] also observes similar disappearing phenomena of disassortative mixing in Czech network excluding prepositions.

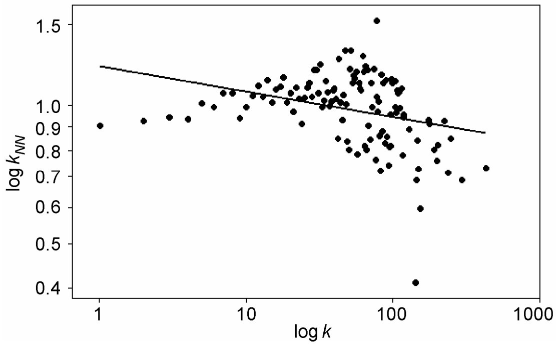

Figure 7 The average nearest-neighbor degree as a function of the node degree. The slope of regression line is –0.055.

3 Discussion

Based on the statistical patterns of the semantic network and the corresponding E-R random network, we therefore conclude that semantic network is scale-free and almost small-world. However, semantic network has smaller clustering coefficient and greater diameter and average path length than that of syntactic networks.

The greatest differences between semantic and syntactic network are: (1) the correlation between the clustering coefficient and the degree of the nodes, (2) the correlation between the degree of a node and that of its neighbors. If we consider that such differences come from the lack of functional words in semantic network, perhaps it is reasonable to consider that semantic (or concept) network has a few different structures from syntactic and other real networks.

In this study, we build and investigate a dynamic semantic network, which reflects semantic structure in practical language usage and processing. There are also a few studies on static semantic networks, such as word associations, WordNet, thesaurus and semantic web[21,22]. Static semantic network is a representation of human (or world) knowledge system or organization.

Ref. [21] presents statistical properties of the large-scale structure of 3 types of semantic networks: word associations, WordNet, and Roget's thesaurus. They show that these networks have a small-world structure and a scale-free pattern of connectivity.

The semantic web is the application of advanced knowledge technologies to the web and distributed systems in general. The core of semantic web technology is ontology-based representation. Ontology is a shared, formal conceptualization of a domain, i.e. a description of concepts and their relationships. Therefore it is possible to build a network based on ontologies. Ref. [22] explores semantic web based on the ontologies at DAML library. The results show that the semantic web has small-worldness and scale-freeness.

Table 2 shows the main statistical properties of these static semantic networks, which are almost small-world and scale-free. However, we cannot find the measure of two correlations mentioned above in these studies on static semantic networks. Therefore we do not know whether static semantic network has also a weaker link than syntactic networks as the dynamic semantic network. It is worthy of further investigation.

Table 2 Main properties of semantic web and other semantic networks[a)]

Network	N	$\langle k\rangle$	C	$\langle d\rangle$	D	γ	C_{rand}	$\langle d_{\mathrm{rand}}\rangle$
Semantic web[22]	56592	4.63	0.152	4.37		1.48	8.95E–05	7.23
Associations[21]	5018	22	0.186	3.04	5	3.01	4.35E–03	3.03
WordNet[21]	122005	1.6	0.0265	10.56	27	3.11	1.29E–04	10.61
Roget[21]	29381	1.7	0.875	5.60	10	3.19	0.613	5.43

a) N, number of nodes; $\langle k\rangle$, average degree; C, clustering coefficient; $\langle d\rangle$, average path length; D, diameter; γ, exponent of power law; C_{rand}, clustering coefficient of random graph; $\langle d_{\mathrm{rand}}\rangle$, average path length of random graph

In summary, although some questions are still open, the findings of the present study are useful to explore whether all language networks have similar statistical properties and other questions put forward in the beginning of this paper. Structurally, semantic network is more similar to conceptual network in the brain. Therefore it is worthy of further research on how to find better statistical patterns to describe linguistic and cognitive universals from the viewpoint of complex networks.

The semantic treebank was annotated by Guan Runchi. Helpful comments come from the anonymous reviewers, Roland Hausser and Li Wenwen.

1 Solé R, Corominas B, Valverde S, et al. Language networks: Their structure, function and evolution. Santa Fe Institute Working Paper, 2005, 05-12-042

2 Ferrer i Cancho R, Solé R V, Köhler R. Patterns in syntactic dependency networks. Phys Rev E, 2004, 69: 051915[doi]

3 Li Y, Wei L X, Li W, et al. Small-world patterns in Chinese phrase networks. Chinese Sci Bull, 2005, 50: 286—288[doi]

4 Liu H. The complexity of Chinese dependency syntactic networks. Physica A, 2008, 387: 3048—3058

5 Liu H, Hu F. What role does syntax play in a language network? Europhys Lett, 2008, 83: 18002[doi]

6 Hudson R A. Language Networks: The New Word Grammar. Oxford: Oxford University Press, 2007

7 Tesnière L. Eléments de la Syntaxe Structurale. Paris: Klincksieck, 1959

8 Hajičová E. Dependency-based underlying-structure tagging of a very large Czech corpus. In: special issue of TAL journal, Grammaires de Dépendence/Dependency Grammars. Paris: Hermes, 2000. 57—78

9 Milićević J. A short guide to the meaning-text linguistic theory. J Koralex, 2006, 8: 187—233

10 Hausser R. A Computational Model of Natural Language Communication. Berlin, Heidelberg: Springer, 2006

11 Sowa J F. Conceptual graphs for a data base interface. IBM J Res Develop, 1976, 20: 336—357

12 Guan R. Annotation of Chinese semantic treebank and semantic parsing (in Chinese). Master Thesis. Beijing: Communication University of China, 2008

13 Albert R, Barabási A L. Statistical mechanics of complex networks. Rev Mod Phys, 2002, 74: 47—97[doi]

14 Watts D J, Strogatz S H. Collective dynamics of "small-world" networks. Nature, 1998, 393: 440—442[doi]

15 Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509—512[doi]

16 Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks. Science, 2002, 297: 1551—1555[doi]

17 Pastor-Satorras R, Vespignani A. Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge, UK: Cambridge University Press, 2004

18 Newman M E J. Assortative mixing in networks. Phys Rev Lett, 2002, 89: 208701[doi]

19 Pastor-Satorras R, Vazquez A, Vespignani A. Dynamical and correlation properties of the internet. Phys Rev Lett, 2001, 87: 258701[doi]

20 Caldarelli G. Scale-free Networks: Complex Webs in Nature and Technology. Oxford: Oxford University Press, 2007

21 Steyvers M, Tenenbaum J B. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognit Sci, 2005, 29: 41—78

22 Gil R, García R. Measuring the semantic web. In: Advances in Metadata Research, Proceedings of MTSR'05. Paramus: Rinton Press, 2006