Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Comparison of co-occurrence networks of the Chinese and English languages

Wei Liang [a], Yuming Shi [a,*], Chi K. Tse [b], Jing Liu [c], Yanli Wang [a], Xunqiang Cui [a]

[a] *Department of Mathematics, Shandong University, Jinan, Shandong 250100, China*

[b] *Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China*

[c] *State Key Laboratory of Software Engineering, Wuhan University, Wuhan, Hubei 430072, China*

ARTICLE INFO

Article history:
Received 26 May 2009
Available online 4 August 2009

PACS:
89.75.Da
89.75.Fb

Keywords:
Language
Co-occurrence network
Small-world network
Scale-free network

ABSTRACT

Co-occurrence networks of Chinese characters and words, and of English words, are constructed from collections of Chinese and English articles, respectively. Four types of collections are considered, namely, essays, novels, popular science articles, and news reports. Statistical parameters of the networks are studied, including diameter, average degree, degree distribution, clustering coefficient, average shortest path length, as well as the number of connected subnetworks. It is found that the character and word networks of each type of article in the Chinese language, and the word network of each type of article in the English language all exhibit scale-free and small-world features. The statistical parameters of these co-occurrence networks are compared within the same language and across the two languages. This study reveals some commonalities and differences between Chinese and English languages, and among the four types of articles in each language from a complex network perspective. In particular, it is shown that expressions in English are briefer than those in Chinese in a certain sense.

1. Introduction

Complex networks have attracted a great deal of interest since the publication of the works of Watts and Strogatz [1] and Barabási and Albert [2]. Recently, complex network theory has been widely used to study some behaviors of complex systems in the real world such as World Wide Web and Internet [3–6], biological networks [7], collaboration networks [8], and public transport networks [9,10]. The use of the complex network approach has been found fruitful in the analysis of a variety of complex systems.

Human languages can be studied in terms of complex network models. Recently, language networks have been constructed with different criteria for connecting words or characters, such as co-occurrence [11–13], syntactic dependency [14], and semantic dependency [15–17]. These networks exhibit the small-world or scale-free feature, or both. The study of language networks has also been applied in language learning and its evolution [18,19], quantification of language characteristics (e.g. Zipf's law [20]), and comparative study among two or more languages [21,22].

There are at least 6800 different languages in the world [23]. The Chinese and English languages are two of the mostly spoken ones. For these two languages, word co-occurrence networks have been widely studied [11–13] and have been shown to exhibit small-world and scale-free features. Sentences in Chinese are formed by characters and words, while sentences in English are formed by words. Character co-occurrence networks can be constructed in a likewise manner as in the construction of word co-occurrence networks. As yet, no study has been performed on the character networks

[*] Corresponding author.
E-mail addresses: liangwei5321@yahoo.com.cn (W. Liang), ymshi@sdu.edu.cn (Y. Shi), encktse@polyu.edu.hk (C.K. Tse), j_liu@whu.edu.cn (J. Liu).

0378-4371/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2009.07.047

except for our conference paper [24]. In the existing literature, study has focused on a single network that is constructed from a large number of articles, which are selected from tagged corpus, wordnet, online English dictionary, etc. However, a character co-occurrence network and a word co-occurrence network can be constructed from a single Chinese article, and a word co-occurrence network can be constructed from a single English article. Do these networks still exhibit small-world and scale-free features? Can useful conclusions be made by comparing network parameters corresponding to two or more languages from a network perspective? In order to answer these questions, we have constructed 114 networks from collections of 53 Chinese articles, including essays, novels, popular science articles, news reports, and 4 concatenated articles of each type [24]. We found that these character and word co-occurrence networks are qualitatively equivalent, i.e., they exhibit small-world and scale-free features.

In this paper, based on our previous work on the Chinese language [24], we further study the commonalities and differences between Chinese and English, and among four types of articles, i.e., essays, novels, popular science articles, and news reports, in each language from a complex network perspective. In order to achieve this goal, 200 Chinese articles and 200 English articles are selected. A character co-occurrence network and a word co-occurrence network are constructed from a single Chinese article, and a word co-occurrence network is constructed from a single English article. Therefore, 400 Chinese character and word co-occurrence networks and 200 English word co-occurrence networks are constructed. Furthermore, in order to confirm the results concluded from analyzing the above 600 networks, 10 articles including 5 essays and 5 novels which have both Chinese and English versions are selected, where the Chinese versions are translated from their corresponding English versions, and their corresponding networks are constructed. All these networks are treated as undirected and unweighted graphs. Their statistical parameters are studied, including diameter, average degree, degree distribution, clustering coefficient, average shortest path length as well as the number of connected subnetworks. They are shown to exhibit scale-free and small-world features. We compare the statistical parameters of these networks in the same language and across the two languages and find that some statistical parameters of co-occurrence networks of different languages and different article types are almost identical while others can be quite different. This study reveals some commonalities and differences between the Chinese and English languages, and among different types of articles in each language from a complex network perspective. In particular, our empirical results show that expressions in English are briefer than those in Chinese in a certain sense.

2. Some basic concepts of complex networks

A *network* G is a set of nodes V with edges E, denoted by $G = (V, E)$. Suppose that a network with N nodes is undirected and unweighted. The *degree* of node i is the number of edges that the node has, denoted by k_i. The average degree of the network is defined by $\langle k\rangle = \sum_{i=1}^{N} k_i/N$. A network is said to be *connected* if for any two nodes in the network there is at least a path to connect these two nodes. Given two nodes $i, j \in V$, let d_{ij} be the shortest path length that connects them. Suppose that the network is connected, the diameter of the network is defined by

$$D = \max_{1\le i,j\le N} d_{ij},$$

and the *average shortest path length* of the network is defined as

$$L = \frac{2\sum\limits_{i>j} d_{ij}}{N(N-1)}.$$

The *clustering coefficient* C_i of node i is the probability that any two neighbors of node i are also connected to each other, i.e.,

$$C_i = \frac{2E_i}{k_i(k_i-1)},$$

where E_i is the number of the actual edges among the neighbors of node i. The clustering coefficient of the whole network is the average of C_i $(1 \le i \le N)$, denoted by C.

A random graph can be obtained by linking pairs of nodes with some probability. One of the most important random graphs is the Erdös–Rényi graph, which has a binomial degree distribution that can be approximated by a Poisson distribution [25]. For an Erdös–Rényi graph with an average degree $\langle k\rangle$, its average shortest path length is $L_r \approx lnN/ln\langle k\rangle$ and its clustering coefficient is $C_r \approx \langle k\rangle/(N-1)$. A network is said to be a *small-world network* if its average shortest path length $L \approx L_r$ and its clustering coefficient $C \gg C_r$.

Degree distribution $p(k)$ is one of the most important statistical characteristics of a network, which is defined as the probability that a randomly chosen node in the network has exactly degree k. If $p(k)$ satisfies the power-law degree distribution:

$$p(k) \propto k^{-\gamma},$$

where γ is a positive constant, then the network is said to be *scale free.*

Remarks: For convenience, we set $d_{ij} = 0$ if there is no connection between nodes i and j of a network, and $C_i = 0$ if there is no connection between node i and other nodes of a network. Then D, L, and C for a non-connected network can be likewise defined.