Physica A 393 (2014) 268–276

Contents lists available at ScienceDirect

Physica A

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

Modeling the Chinese language as an evolving network

Wei Liang [a], Yuming Shi [b,*], Qiuling Huang [c]

[a] School of Mathematics and Information Science, Henan Polytechnic University Jiaozuo, Henan 454000, China
[b] Department of Mathematics, Shandong University Jinan, Shandong 250100, China
[c] School of Mathematics and Quantitative economics, Shandong University of Finance and Economics Jinan, Shandong 250014, China

HIGHLIGHTS

- An evolving language network model is proposed.
- The model includes adding nodes and edges as well as rewiring and deleting edges.
- Interesting phenomena are found in fitting the networks in 11 different periods of China.

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form 22 July 2013
Available online 3 September 2013

Keywords:
Language
Model
Scale-free
Character co-occurrence network

ABSTRACT

The evolution of Chinese language has three main features: the total number of characters is gradually increasing, new words are generated in the existing characters, and some old words are no longer used in daily-life language. Based on the features, we propose an evolving language network model. Finally, we use this model to simulate the character co-occurrence networks (nodes are characters, and two characters are connected by an edge if they are adjacent to each other) constructed from essays in 11 different periods of China, and find that characters that appear with high frequency in old words are likely to be reused when new words are formed.

1. Introduction

Language is the quintessence of human civilization. It is a complex adaptive system that has evolved through the ages [1]. Chinese language networks have been wildly studied in the recent years, such as co-occurrence, syntactic dependency, and semantic dependency [2–9]. These networks exhibit either the small-world or scale-free feature, or both. Chinese history and culture have experienced a long time of development. What are similarities and differences in Chinese language in the different Chinese historical periods from a network perspective? Does the evolution of Chinese language have a certain rule? To answer these questions, we constructed 561 networks (nodes are characters, and two characters are connected by an edge if they are adjacent to each other) from 550 essays in 11 different periods and 11 concatenated articles of each period [9]. We found that 99.6% of the networks have the scale-free feature, and 95.0% have the small-world feature. This provides some necessary statistical data to establish a model of the evolution of Chinese language.

Recently, there are a number of papers utilizing the Google n-gram data to analyze the statistical properties of written language. For example, Petersen et al. identified tipping points in the life trajectory of new words, statistical patterns, and quantitative measures for cultural memory [10]; Gao et al. found that words describing social phenomena tend to have different long-range correlations than words describing natural phenomena [11]; Petersen et al. analyzed the rates of the appearance of new words and the disappearance of old words in language [12]; Perc studied the evolution of the most common English words and phrases over the centuries [13].

* Corresponding author.
E-mail addresses: wliang@hpu.edu.cn (W. Liang), ymshi@sdu.edu.cn (Y. Shi), hql_shj@163.com (Q. Huang).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.08.034

In order to explain the mechanism of causing a power-law degree distribution, Barabási and Albert built the BA model with the power-law exponent, γ, is 3 [14]. However, γ is scattered between 1 and 4 in most real-world networks [15,16]. Albert and Barabási built another model in which m new edges are added with probability p, m edges are rewired with probability q, and a new node with m edges is added with probability $1-p-q$ [17]. They proposed a continuum theory to predict the degree distribution of the model, and found that if

$$q < \min\{1-p, (m+1-p)/(2m+1)\},$$

then it is scale free and γ varies from 2 to 4, and if $q \to 1$, then the model develops an exponential tail [17]. Shi et al. built a model in which a new node with m edges are added as well as c links are removed, and found that the network is scale free with γ varying from 1 to 4 if $m \geq c$ [18].

There are a few papers in the study of modeling language evolving networks. Dorogovtsev and Mendes built a DM model to analyze the degree distributions of two English word co-occurrence networks in Ref. [19]. The DM model was obtained by adding ct edges at time t on the basis of the BA model, and found that $\gamma = 3$ in the region of the kernel lexicon that contains about 5000 words and is the most important core part of language, and $\gamma = 1.5$ in the region of the other lexicon [20]. In order to better simulate the degree distribution in Ref. [19], Markošová built a model by adding rewire edges on the basis of the DM model [21]. There is a quite difference between Chinese and English languages. Chinese words are made of characters just as English words are made of letters; individual Chinese characters each have a meaning, while English letters have no intrinsic meaning. Networks were constructed from the inclusion relationship of Chinese characters or phases by Yu et al., and a model including increasing nodes and preferential attachment was built [22].

Chinese language goes through more than 5000 years development. Based on our data in the study of its evolution in 11 different periods [9], there exists an important feature of its evolution: that is, some old expressions or words are no longer used in daily-life language. This means that connections between some two characters would disappear with evolution. So there should be added the operation: deleting edges in modeling language evolving networks in addition to the operations: increasing nodes and rewiring edges with preferential attachment. However, this is not considered in the above three models. In order to characterize the Chinese language evolution, we build a model of a Chinese character network in which a new node is added, and edges are added, rewired, and deleted in the present paper. We calculate the degree distribution of the model, and find that the degree distribution is power law in some case, where γ is scattered between 1 and $+\infty$, and the degree distribution is exponential in some other case. The parameters of the model for simulation are determined by the practical statistical parameters of the networks that we obtained in Ref. [9]. We find that when a new word or expression is formed in the Chinese language evolution, the selection of characters has strong preference, that is, characters that appear with high frequency in old words are likely to be reused when new words are formed. In addition, all the above existing models in have not simulated networks constructed from different periods of English or Chinese languages.

2. Evolution of Chinese language

Chinese language has experienced a long time of development. The evolution of Chinese language has the following features:

(1) The number of characters is gradually increasing. How many characters were there in different periods? According to statistics, the number of Oracle discovered so far is about 4500 [23]. Qin Shi Huang (259 BC–210 BC) unified the six countries and their cultures, and pushed the rapid development of characters. In the Han Dynasty (206 BC–220), the number of characters was significantly increased. "Shuo Wen Jie Zi" (in Chinese "说文解字") is the earliest Chinese dictionary, which contains 9353 characters. In the Three Kingdoms period (220–280), "Sheng Lei" (in Chinese "声类"), written by Deng Li, includes 11,520 characters, and "Guang Ya" (in Chinese "广雅" contains 18,151 characters. In the Southern–Northern Dynasties (420–589), "Yu Pian" (in Chinese "玉篇") contains 22,726 characters [24]. In the Song Dynasty (960–1279), "Guang Yun" (in Chinese "广韵") contains 26,194 characters, and "Lei Pian" (in Chinese "类篇") contains 33,190 characters [25]. "Zheng Zi Tong" (in Chinese "正字通"), written by Zilie Zhang, contains 33,440 characters in the Ming Dynasty (1368–1644). There are 87019 characters in "Kangxi Dictionary" (in Chinese "康熙词典"), written by Yushu Zhang, in the Qing Dynasty (1644–1911). In Beijing Guoan Consulting Equipment Company, the number of characters that have been identified by experts is 91,251, which consists of the most comprehensive characters.

(2) New words are continuously generated in the existing characters. For example, 15,400 new words have been created since the People's Republic of China was established, and they are included in "Contemporary Chinese New Words Dictionary" [26].

(3) Some old expressions or words are no longer used in daily-life language. For example, "red guards" (in Chinese "红卫兵"), which means a special group composed of young students, and was created during the Chinese Cultural Revolution (1966–1976).

Therefore, adding nodes and edges, and rewiring and deleting edges should be included when one builds a model to characterize the evolution of Chinese language. Moreover, in the generation of a new edge, despite the nodes with high degrees tend attract more new connections, there is also random selection. In fact, we have known that the nodes with large number of connections are "的", "了", and "是" in the modern Chinese language networks [8]. When some new words, such as "菜鸟", "美眉", were created [3,26], the selection of characters is of randomness.