Physica A 412 (2014) 84–91

Contents lists available at ScienceDirect

Physica A

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

Structure and modeling of the network of two-Chinese-character compound words in the Japanese language

Ken Yamamoto [a,*], Yoshihiro Yamazaki [b]

[a] Department of Physics, Faculty of Science and Engineering, Chuo University, Kasuga, Bunkyo-ku, Tokyo 112-8851, Japan
[b] Department of Physics, School of Advanced Science and Engineering, Waseda University, Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

HIGHLIGHTS

- We have measured two-character compounds of the Japanese language by using complex network analysis.
- We have applied the HITS algorithm and PageRank and calculated their correlation to in- and out-degrees.
- We have proposed a numerical model based on the property that an important character tends to get many edges.
- We have successively made a quantitative comparison between the proposed model and an actual network.
- We have discussed the relevance of our model to the fitness model.

ARTICLE INFO

Article history:
Received 6 September 2013
Received in revised form 31 March 2014
Available online 25 June 2014

Keywords:
Complex network
Chinese character
Japanese language
Fitness model
Directed analyses

ABSTRACT

This paper proposes a numerical model of the network of two-Chinese-character compound words (two-character network, for short). In this network, a Chinese character is a node and a two-Chinese-character compound word links two nodes. The basic framework of the model is that an important character gets many edges. As the importance of a character, we use the frequency of each character appearing in publications. The direction of edge is given according to a random number assigned to nodes. The network generated by the model is small-world and scale-free, and reproduces statistical properties in the actual two-character network quantitatively.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Research fields of network science are increasingly spreading, and complex network analysis currently has become a fundamental piece in understanding complex systems. Network analysis in physics dates back to the discoveries of the small-world [1] and scale-free [2] properties. Small-world means the coexistence of small path length and high clustering, and scale-free means the power law of degree distribution: $P(k) \propto k^{-\gamma}$. Some other features, such as the community structure [3], network motif [4], and hierarchical structure [5], have been developed.

Human languages are considered to be typical complex systems, whose structures have been described by complex networks from various aspects: thesaurus networks [6], word association networks [7], word co-occurrence networks [8], syntactic dependency networks [9], and so on. Knowledge of language networks has been also applied to text mining [10], natural language processing [11], and language acquisition [12].

* Corresponding author. Tel.: +81 3 3817 3374.
E-mail address: yamamoto@phys.chuo-u.ac.jp (K. Yamamoto).

http://dx.doi.org/10.1016/j.physa.2014.06.031
0378-4371/© 2014 Elsevier B.V. All rights reserved.

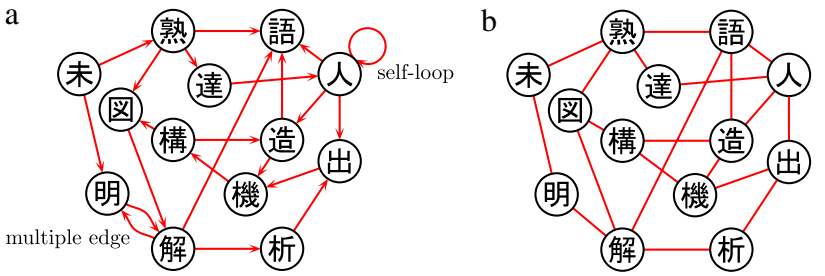

Fig. 1. A small part of the two-character network of the *Kojien* dictionary, which is a directed network having multiple edges and self-loops (a). The undirected counterpart (b) is also considered in this paper.

Table 1
Fundamental characteristics of the three dictionaries. The average path length l is small, and the clustering coefficient C is much larger than C_{random} which is the average clustering coefficient of a random network having the same number of nodes and edges as the actual network. The power-law exponent γ of the degree distribution [$P(k) \propto k^{-\gamma}$] is also shown.

Dictionary	Nodes	Edges	$\langle k \rangle$	l	C	C_{random}	γ
Kojien	5458	74 617	27.3	3.14	0.138	0.00501	1.14
Iwanami	3904	32 150	16.5	3.31	0.085	0.00424	1.16
Sanseido	3444	28 358	16.5	3.32	0.086	0.00483	1.14

The present authors have studied the network structure of Chinese characters in the Japanese language [13]. The Japanese language has many words composed of two Chinese characters (these words are called *niji-jukugo* in Japanese). For example, the characters " 漢 " (Han or China) and " 字 " (character) form the compound " 漢字 " (Chinese character). It was reported that the two-character words account for 70% of the headwords of a Japanese-language dictionary [14]. We constructed the *two-character network* by regarding each two-character compound as an edge connecting two Chinese characters (nodes). Fig. 1(a) is a part of this network. The first and second characters in a compound cannot be inverted generally, so the network is directed; the edge direction is indicated by an arrow from the first character to the second. A few compounds can be invertible [" 解明 " (clarify) and " 明解 " (clear and lucid) in Fig. 1(a) for example], and they form multiple edges and self-loops in a two-character network. From the analysis of the undirected counterpart as in Fig. 1(b), we have previously found that networks built from headwords of the three Japanese-language dictionaries, *Kojien*, *Iwanami Kokugo Jiten*, and *Sanseido Kokugo Jiten* [15], are small-world and scale-free in common (Table 1). A similar study was carried out in the Chinese language [16]. Small-world and scale-free properties are confirmed also in Chinese networks, but their scale-free exponents $\gamma = 1.40$ (Standard Chinese) and 1.49 (Cantonese) are different from those of the Japanese networks.

The aim of this paper is to propose a stochastic model for reproducing the statistical properties of the two-character network. A basic point of the model is that a Chinese character easily gets edges when it possesses high importance. This model generates a directed network. However, directed network analysis of two-character networks has not been done sufficiently in previous studies [13,16], so we give directed network analyses of the HITS (hyperlink-induced topic search) algorithm and PageRank for a two-character network before the explanation of the model. We confirm that the proposed model is quantitatively consistent with the actual two-character network of *Kojien*.

2. Directed network analysis: HITS and PageRank

In this section, we apply the HITS algorithm and PageRank to the *Kojien* network. These two methods were originally developed to rate the importance of web pages based on hyperlinks, and are now employed widely in analyses of many directed networks.

The HITS algorithm assigns hub and authority scores to each node [17]; a node gets a high hub score if it has many outgoing edges to nodes with high authority scores, and a node gets a high authority score if it has many incoming edges from nodes with high hub scores. Fig. 2 shows that the authority score highly correlates with indegree (Panel (a)), and that the hub score highly correlates with outdegree (Panel (b)).

Along with the HITS algorithm, the PageRank is also a technique to assign the importance of each node in a directed network [18]. The idea of the PageRank is based on a relation that a web page which receives many hyperlinks from many important pages is a good page. From the viewpoint of statistical physics, the PageRank is a kind of the visiting probability of a random walk which hops along the edge direction. Technically, the random walker teleports with probability q to an arbitrary node so that the walker is not trapped in a node having no outgoing edges. (The term "teleport" is officially used in research works of the PageRank.) According to a study of the PageRank [19], people follow six hyperlinks on average until they begin a new search, and hence an appropriate teleport probability is $q \approx 1/6$. In fact, the standard value is $q = 0.15$ in research works of web networks [18–20]. However, there is no clear advantage in considering a long sequence of nodes in our two-character network, because one does not wander from character to character. Only the structure of whether two

Download English Version:

https://daneshyari.com/en/article/7380367

Download Persian Version:

https://daneshyari.com/article/7380367

Daneshyari.com