Physica A 415 (2014) 307–314

Contents lists available at ScienceDirect

Physica A

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

Complex dynamics of text analysis

Xiaohua Ke [a,b,*], Yongqiang Zeng [b,c], Qinghua Ma [a], Lin Zhu [a]

[a] Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, China
[b] Social Science Key Laboratory of Language Engineering and Computing of Guangdong Province, Guangzhou, China
[c] Guangdong Teachers College of Foreign Language and Arts, Guangzhou, China

HIGHLIGHTS

- We examine changes in the level of text quality using complex network measurements like degrees, clustering conferences, and dynamical networks.
- Complex network features of different text qualities can be clearly revealed.
- Decreasing amounts of nodes in complex networks indicating higher quality of writing.
- Complex network theories can be applied to potential applications of text analysis.

ARTICLE INFO

Article history:
Received 29 March 2014
Received in revised form 12 July 2014
Available online 19 August 2014

Keywords:
Complex network
Text analysis
Text quality series variation
Network feature

ABSTRACT

This paper presents a novel method for the analysis of nonlinear text quality in Chinese language. Texts produced by university students in China were represented as scale-free networks (word adjacency model), from which typical network features such as the in/outdegree, clustering coefficient and network dynamics were obtained. The method integrates the classical concepts of network feature representation and text quality series variation. The analytical and numerical scheme leads to a parameter space representation that constitutes a valid alternative to represent the network features. The results reveal that complex network features of different text qualities can be clearly revealed and applied to potential applications in other instances of text analysis.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

In China, achievement tests[1] are common for measuring the effects that learning and teaching have on individuals. In college, they were adopted in most courses via short writings. Text quality plays a fundamental role in assessing writing activities whereas text features of score series reveal complex dynamical phenomena. For the last few years, nonlinear dynamics of complex network have been employed as statistical or stochastic analytical methods of measurement in financial, electrical, mechanical, thermal, and other types of physical systems [1]. Because the representation and analysis of written texts in terms of complex networks offer a promising and challenging research opportunity for autoscoring writings and translations [2,3], it is worthy of involving mathematical tools usual in the analysis of nonlinear dynamics from text quality series variation.

* Correspondence to: 2# Baiyun Dadao, Guangzhou, China.
E-mail addresses: Ck0900@hotmail.com (X. Ke), 199110513@oamail.gdufs.edu.cn (Y. Zeng), gdqhma@21cn.com (Q. Ma), 1107509010@qq.com (L. Zhu).

[1] An achievement test is to measure skills and knowledge learned in a given grade level. Its scores are often used in an educational system to determine what grade-level a student belonging to.

http://dx.doi.org/10.1016/j.physa.2014.08.022
0378-4371/© 2014 Elsevier B.V. All rights reserved.

By using unilingual or bilingual corpus, some researchers managed to adopt complex networks to investigate texts' formats and content by several measurements, like the immediate associations between words, and the node degree, clustering coefficient and network dynamics [4]. Other researchers used complex network theory to retrieve information on some properties of the text, such as style and authorship [5], measure the relevance of words in a text [2], and identified the works of renowned English writers from the perspective of first-order statistics of words and other features obtained from texts [6]. A series of studies indicated that word adjacency networks [7], word association networks [8], semantic networks [1], and syntactic networks [9] are graphs that show features present in classical examples of complex networks. One of the important consequences of such studies is the presence of hubs in linguistic networks, e.g. the World Wide Web and social networks.

This paper presents a novel method for the analysis of nonlinear text quality in Chinese language. In Section 2, a method based on complex network theory is presented to analyzes text features like structure, coherence and cohesion, adherence to writing topics, and theme adequacy. Several complex networks of texts are constructed based on different score levels. In Section 2.3, we provide the discussions, conclusions, and future work.

2. Modeling approach and experiments

In the procedure for constructing complex networks of language, we firstly collect over 500 short-answer responses to a topic-discussing prompt. It was in a final exam that all the students read the prompt and write a response in Chinese language with an average article size of 173 words (or 205 characters). For better understanding, we translated several response samples, several key words into English in this paper, although they were originally written and processed in Chinese. At the same time, we provided the prompt and reference answer, and selected two samples of short-answer responses to be displayed in the Appendix.

To meet the challenge of maintaining testing reliability, each text was firstly scored by two human raters, in accordance with holistic rubrics that described six levels of performance, scored 1 through 6, where a score of "6" indicates the *highest* quality writing, and a score of "1" indicates one of the *lowest* quality, otherwise 0 was used to designate 'unresponsive' short-writings (texts that are off topic, repetitions of the prompt, or otherwise unintelligible or blank). Most large-scale tests require essays to be double-rated for reliability, with a third rater used if the first two raters disagree [9,10]. In our research, if the scores from the two human raters differed by less than two points, the average of the two scores became the final score. If the first two readers disagreed by two points or more, the essay was then rated by a third human rater and the final score was the average of the two closest scores. Human raters used 2 criteria related to text quality, namely (i) adherence to standard short-writing sample and (ii) theme adequacy/development. These are the criteria generally employed to mark short-writings for university students applying to achievement tests in China. Because of the large score dispersion for some texts, we select 80 texts within scores of 3, 4, 5, and 6, which means there are totally 320 texts in this research.

Then, we construct word adjacency/co-occurrence complex networks by considering the proximity between words. In the first step, we do the pretreatments for all the text, including: (i) Separate each texts into sentences ($S_1, S_2, \ldots S_n$). (ii) For every sentence, a Chinese word segmentation program[2] had been launched to obtain separated words, e.g. $w_1; w_2; w_3; \ldots$, implying a directed edge from w_1 to w_3 and another directed edge from w_2 to w_3. (iii) The stopwords are removed and most of the synonym words were merged. Stopwords are also known as non-meaningful words, such as verb to be, some adverbs, and words from closed classes like articles, pronouns, prepositions and conjunctions. Then, the remaining words were merged according to a synonym corpus. Therefore, different occurrences of meaning-related words are represented uniquely by the same node in the complex network. In the second step, word adjacency networks of different score-levels were obtained. Obviously, there are 80 texts in a score-level, e.g. texts with a score of 3, and each word in a text can be represented as a node, while subsequent words are defined associations. Nodes are connected with edges when they appear in one sentence of the text. Typically, pairs of subsequent words, excluding texts and other connecting words are connected with unity length by considering couples of subsequent words.

Thirdly, three measurements of complex networks are estimated after the construction of the networks. That means the indegree and outdegree, the clustering coefficient, and the network dynamics are estimated while the number of connected components is monitored during complex networks' growing. So as to calculate topological network features considering word associations in a complex network. Keeping in mind that each complex network had been stored as one weight matrix, for example matrix A, all network measurements calculated here were performed based on the specific matrix.

2.1. Indegree and outdegree

Each node i in the complex network has its own indegree and outdegree. The in/outdegrees are defined respectively as

$$ID_i = \sum_{j=1}^{N} A_{ij}, \tag{1}$$

[2] We use the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System).

Download English Version:

https://daneshyari.com/en/article/7380105

Download Persian Version:

https://daneshyari.com/article/7380105

Daneshyari.com