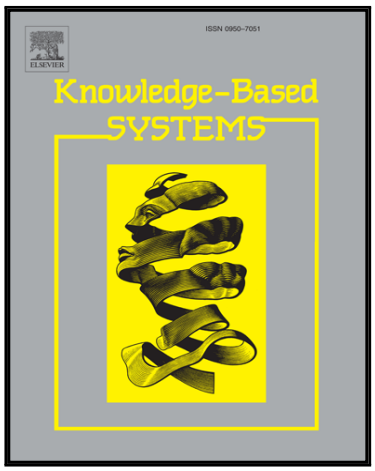

Long range dependence in texts: A Method for quantifying coherence of text

Elham Najafi, Amir H. Darooneh

PII: S0950-7051(17)30312-X
DOI: 10.1016/j.knosys.2017.06.032
Reference: KNOSYS 3960

To appear in: *Knowledge-Based Systems*

Received date: 12 March 2017
Revised date: 21 May 2017
Accepted date: 25 June 2017

Please cite this article as: Elham Najafi, Amir H. Darooneh, Long range dependence in texts: A Method for quantifying coherence of text, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.06.032

Long range dependence in texts: A Method for quantifying coherence of text

Elham Najafi[a,*], Amir H. Darooneh[a]

[a]*Department of Physics, University of Zanjan, P.O.Box 45196-313, Zanjan, Iran*

Abstract

This paper discusses a major issue in computational linguistics; the automatic calculation of text coherence. Heretofore, only few methods have been proposed to automatically detect local coherence of texts. All of these methods need a lot of pre-processing tasks and computational efforts. Here we suggest a simple method to evaluate the coherence globally. First, we use a word ranking method to assign an importance value to each word-type in a text, then the importance time series associated with text is constructed. In the next step, Detrended Fluctuation Analysis(DFA) which is used for detecting inherent correlations in time series, is applied to texts importance time series. We found that the importance time series exhibits a bi-scale behavior; it is long-range correlated at large distances, while short-range correlations are observed in small distances. We also observed that for a shuffled text the scaling exponent decreases. This decrease becomes more and more significant when we reshuffle the chapters, paragraphs, sentences and words respectively. This fact leads us to consider the scaling exponent of text time series (or briefly STT) as a measure for quantifying the global coherence. We demonstrate our claim by carrying out an experiment on three sample texts and comparing our method by some entity grid based models.

Keywords: Global Coherence, Importance time series, Scaling Exponent

1. Introduction

Much of the human cognition can be interpreted as a mental process for changing the sensory perceptions into coherent patterns of concepts. This coherence demonstrates itself in the ordering of words when one tries to communicate his ideas by writing or speaking. Quantifying the coherence in texts may help us to explore some hidden aspects of cognitive processes [1, 2, 3, 4, 5]. Automatic coherence evaluation is also an interesting subject in artificial intelligence, linguistics, data mining and some other disciplines.

Coherence is a quality of written or spoken texts that makes them easy to read and understand [15, 16]. A coherent text obeys a particular logical order and is easy to comprehend as a unit, instead of a bunch of messy sentences. Text coherence occurs in two levels; local and global. Local coherence is representative of similarity among adjacent parts of a text, for example, adjacent paragraphs or sentences. On the other hand, global coherence is indicator of the connection between all segments of a text as a whole [17].

Text is an ordered sequence of words that can be considered as an one dimensional discrete space. The meaning of text regulates the distribution of words throughout this space. All word types have self similar distributions across the text but with different fractal features. An importance value can be assigned to each word type by considering its fractal pattern [6]. We associate a time series to every text by substituting the words with their corresponding importance values while retaining their order. The importance time series exhibits the long range dependence for meaningful texts. We assert in this work that the long range correlation in the importance time series is related to the global coherence of the text.

The detrended fluctuation analysis (DFA) is the most powerful method for exploring the long-range correlations in non-stationary time series [8, 9, 10]. We use several variants of this method for analyzing three sample texts. We find that importance time series of texts displays bi-scale behavior; There is short range correlation between words at small

*corresponding author

Email addresses: e.najafi@znu.ac.ir (Elham Najafi), darooneh@znu.ac.ir (Amir H. Darooneh)