Neurocomputing 147 (2015) 147–159

Contents lists available at ScienceDirect

Neurocomputing

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

Stylistics analysis and authorship attribution algorithms based on self-organizing maps

Antonio Neme [a,b,*,1], J.R.G. Pulido [c], Abril Muñoz [d], Sergio Hernández [e], Teresa Dey [f]

[a] Complex Systems Group, Universidad Autónoma de la Ciudad de México, San Lorenzo 290, México, D.F., Mexico
[b] Institute for Molecular Medicine Finland, Tukholmankatu 5, 00270 Helsinki, Finland
[c] Faculty of Telematics, Universidad de Colima, Mexico
[d] CINVESTAV IDS, México D.F., Mexico
[e] Postgraduate Program in Complex Systems, Universidad Autónoma de la Ciudad de México, Mexico
[f] Faculty of Literary Creation, Universidad Autónoma de la Ciudad de México, Mexico

ARTICLE INFO

Article history:
Received 14 April 2013
Received in revised form
27 February 2014
Accepted 3 March 2014
Available online 27 June 2014

Keywords:
Computational stylistics
Authorship attribution
Self-organizing maps
Anomaly detection
Feature selection

ABSTRACT

The style followed by authors can be thought of as a collection of attributes that defines the stylistics space. Texts from the same author tend to be similar in that space. However, the identification of stylistics spaces has proven to be challenging. Associated with the stylistics space is the authorship attribution task. On it, a text of unknown authorship is presented to a system, and the system is expected to identify the author of the text. Two modules define an authorship attribution algorithm: the stylistics space and a classifier. We present a methodology that includes both, a module that allows the identification of novel stylistics spaces, and a classifier to confront the authorship attribution task from the features that define space. The methodology imbricates feature selection, anomaly detection, classification, and visualization algorithms. We applied the capabilities of self-organizing maps not only for visualization but also for anomaly detection, which defines the basis of the classifier. We compared our authorship attribution algorithm with two existing ones. Our methodology achieved similar or better results under *bag-of-words*-related stylistics spaces, and it presented the lowest error under a novel stylistics space based on the rate of introduction of new words.

1. Introduction

The style followed by authors in their texts has been subject to a wide variety of studies from several perspectives [1]. The style has been associated with the use of certain words, the avoidance of others, a bias towards some grammatical structures, or any other measurable pattern [2]. The study of the style is generally known as stylistics [3].

The term *authorship attribution* (AA) is closely related to stylistics. AA refers to the task of identifying the author of a text from a group of candidate authors, based on some possible relevant features extracted from the text [1]. Stylistics may be thought of as the identification of the relevant attributes that define a high-dimensional space in which authors can be recognized from each other. Writers use language following different ways to express their ideas. This variation in language allows authorship attribution to be possible [4].

The motivation to understand the style followed by an author is wide. Among them, the academic reasons are of the highest interest. The study of the literary aspects includes the analysis of the evolution followed by authors along their careers, and the comparison of the style of two authors. Finally, the identification of patterns in texts from a given author is within the interests of literary theory [5]. Several other features relevant to understand the literary aspects can be found elsewhere [2].

A second field in which computational stylistics has an impact is in forensic linguistics. Here, the authorship of a text has to be attributed either for historical reasons or for criminal aspects [1,6]. Also, as a result of high dissemination of thousands of texts in the Web, there are several situations in which it is relevant to assign the authorship of a text. Plagiarism can be detected if similarity in style is detected.

The impact of computational stylistics can also be found in psychiatry. It has been reported that some early symptoms in certain mental disorders can already be detected in writing [7]. For example, a detailed analysis over the novels of Iris Murdoch shows

* Corresponding author at: Complex Systems Group, Universidad Autónoma de la Ciudad de México, San Lorenzo 290, México, D.F., Mexico.
E-mail addresses: antonio.nemecastillo@uef.fi, neme@nolineal.org.mx (A. Neme), jrgp@ucol.mx (J.R.G. Pulido), sergiohzlz.@ciencias.unam.mx (S. Hernández).
[1] Now at the Institute of Biomedicine, School of Health, University of Eastern Finland, Finland.

http://dx.doi.org/10.1016/j.neucom.2014.03.064
0925-2312/

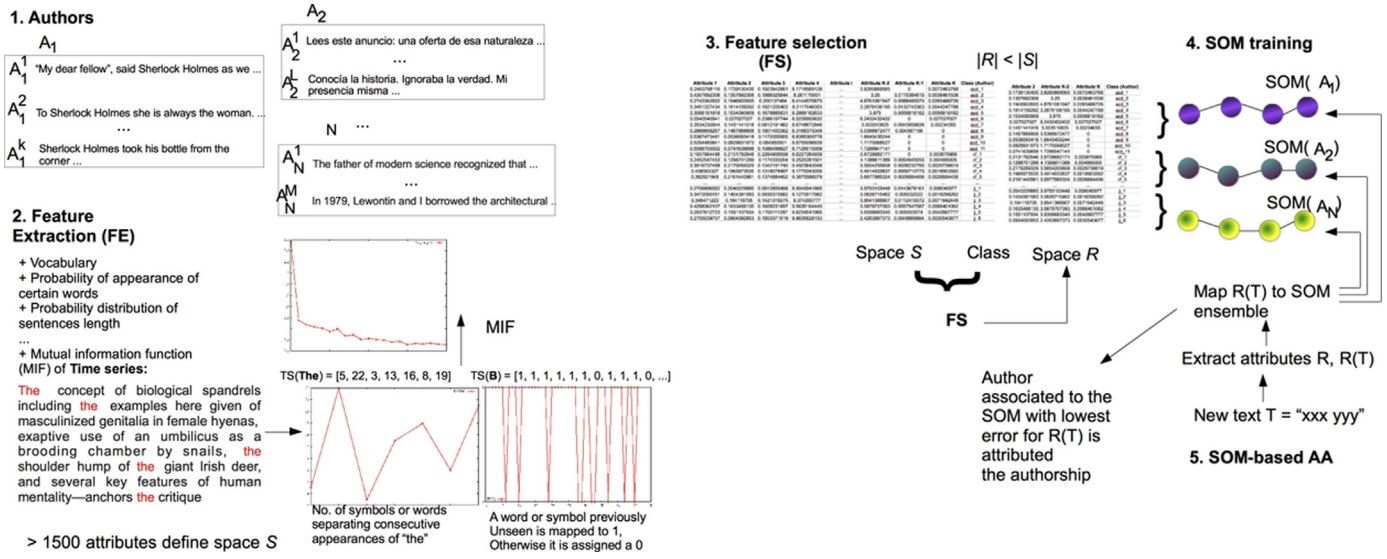

Fig. 1. The proposed methodology for stylistics spaces identification and for AA. (1) For each author A_i, a collection of texts A_i^j are included. (2) Several attributes are extracted from the relevant texts. (3) A feature selection algorithm is applied to the attributes obtained in step 2. (4) An ensemble of SOMs is trained: each SOM is presented only with data from the same author. (5) When a text of unknown authorship arrives, the relevant attributes are extracted from it, and mapped to the SOM ensemble. The author associated with the SOM having the lowest error is attributed the authorship.

that there are qualitative and quantitative differences in her novels prior to the disease and in the early stages of it. Being aware of the general patterns of evolution in stylistics may help psychiatrists and other mental health professionals to early detect symptoms of mental disorders.

In the last two decades, several contributions have defined the field of computational stylistics [1], as an alternative to the human expert in text and style analysis. In a seminal work, Joula [5] presented the AA task within a computational frame. AA is an instance of classification: given an input vector describing a text of unknown authorship, find its class or author. For that, the system, either human or computer, has to be trained in the way authors make use of the relevant features. Two stages are imbricated in AA: (1) the identification of the relevant attributes. (2) The identification of a function that takes those attributes as arguments to classify authors.

Hundreds of attributes have been proposed as relevant for the AA task. Among them, the Burrows's Delta [8] is one of the most clearly defined. It is based on the probability distribution of the 150 most common words in English. Most of the AA algorithms are based on the class of attributes known as *bag of words* (BoW). In the *BoW* approach, texts are represented by the frequency of words, disregarding all relation to grammar [5]. Algorithms based on *BoW* perspectives make the assumption that the style of an author is basically described by the probability distribution of appearance of certain words, phrases, or any other relevant structure [9].

A visualization tool is needed in order to inspect the high-dimensional data describing the studied texts. Among the visualization tools, self-organizing maps (SOM) have been widely applied for visualizing high-dimensional data in general, and for vectors representing texts [10,11] in particular. Also related to the high-dimensionality of stylistics spaces is the task of identifying a subset of attributes that can lead to the identification of authors [2,12]. Feature selection (FS) algorithms are of main relevance at this stage.

We are interested in extending the definitions of stylistics spaces by expanding previous ideas about the benefits of transforming texts into time series [13,4]. Analyzing texts as time series for solving the AA task is an alternative to the more traditional stylistics based on *BoW*. Once texts are transformed into time series, several tools can be applied aiming to detect subtle patterns that can be further investigated as possible discerning attributes.

Our methodology consists of three parts. In the first one, several attributes, including time series representing certain writing patterns, are extracted from texts from different authors (see Fig. 1). The second step consists of the application of a feature selection algorithm to those attributes. The third stage consists of the training of a classifier, based on the attributes selected in stage 2. When a text of an unknown authorship is presented to the trained classifier, it should be able to identify the author. Our classifier is an ensemble of SOMs that operates as an anomaly detector.

The rest of this contribution is presented as follows. In Section 2 we briefly present the related work. In Section 3 we describe the methodology proposed for the study of stylistics and AA. In Section 4 we present the results of applying the proposed methodology to more than 140 texts from 14 professional writers. Finally, we present some conclusions in Section 5.

2. Related work

Open problems and applications of computational author analysis are diverse. In this contribution we describe a novel methodology to face some of the open problems, in particular that of authorship attribution. Telling book authors apart is a well studied problem. In [14], complex networks theory is used to tackle this. Attribution probability is defined based on some measurable quantities of word co-occurrence network of each book. Another classic example of this problem is that of the disputed Federalist papers. Alexander Hamilton and James Madison claimed to have written twelve of those papers [15,16]. With ideas coming from statistical mechanics and information theory, Basile [17] uses similarity measures between pairs of texts as an indication of stylistic closeness. They compare the frequencies of fixed length substrings and compression algorithms as approximation to relative entropy.

A specific type of authorship attribution task, relevant for cybercrime forensic investigations, based on stylometric and