Biologically Inspired Cognitive Architectures (2014) 7, 39–52

Available at www.sciencedirect.com

ScienceDirect

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

RESEARCH ARTICLE

Effects of text essay quality on readers' working memory by a computational model

J. Ignacio Serrano [a,*], M. Dolores del Castillo [b], Ángel Iglesias [a]

[a] Bioengineering Group, Centro de Automática y Robótica, Consejo Superior de Investigaciones Científicas (CSIC), Spain
[b] Bioengineering Group, Consejo Superior de Investigaciones Científicas (CSIC), Spain

Received 16 May 2013; received in revised form 23 October 2013; accepted 23 October 2013

KEYWORDS
Computational modeling;
Reading;
Working memory;
Discourse quality

Abstract
Assessment of essay quality, also called essay scoring, is a task that has been always carried out by human graders. Graders are usually asked to give their scores according to several determined linguistic/semantic criteria. These criteria are related to lexical, syntactical, semantical and discourse features of the texts. In order to replace human graders, automated essay scoring systems make use of statistics on the latter features in order to quantify the quality of the essays. However, there is a subjective component within the evaluation of the text quality that cannot be measured by artificial scorers. Text essays are a form of natural language communication and therefore they cause effects on readers and their cognitive functions. In the work presented in this paper, the dynamic effects that a read text causes on the working memory of readers are studied by means of a connectionist model of memory during reading. Besides, the correlation of those effects with the essay quality scores and text linguistic features is also analyzed. The biologically inspired model of memory includes mechanisms for emulating bounded cognition, getting a little closer to the BICA Challenge achievement. The results obtained also prove how BICA models can feedback Neuroscience and Psychology, thus closing the interdisciplinary loop.

Introduction

Paradoxically, *Natural* Language is *artificially* described by a set of rules worldwide (Pinker, 2000). Human beings are commonly taught to properly use language by following that set of rules. This way, the quality of a language expression or passage can be measured by contrasting it with the corresponding normative description of the language. Thus,

* Corresponding author. Address: Centro de Automática y Robótica, Consejo Superior de Invesigaciones Científicas (CSIC), Ctra. Campo Real km 0.200, La Poveda, 28500 Arganda del Rey, Spain. Tel.: +34 918711900; fax: +34 918717050.
E-mail address: jignacio.serrano@csic.es (J.I. Serrano).

2212-683X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bica.2013.10.002

several linguistic features belonging to different language levels (lexical, syntax, semantics, discourse, topic, etc.) are frequently used to characterize language units (Wang & Brown, 2007) and make comparisons with normative rules.

Nonetheless, Natural Language is a capacity developed as a product of evolution, acquired for the main purpose of communicating with other subjects (this is rather a matter for anthropologists (Duranti, 1997)) with the intention to cause some effects on their feelings, thoughts and ultimately their behavior (Pinker, 1995). From this point, language processing can be considered as a form of coding/decoding (emitter/receptor) of intentions and thoughts into phonemes and graphemes. Consequently, human beings have developed this processing ability and, as a mind ability, it requires cognitive processing and resources.

In turn, Natural Language is a dynamic entity in constant evolution (Christiansen & Kirby, 2003). This evolution of Natural Language has always favored the language use and structures that facilitate communication. There are several factors that can make communication easier, and one of them lies in decoding simplicity and requirements: the less cognitive processing and resources required for decoding the better the understanding. From this statement, a good quality coding implies a soft and easy decoding process (coding means here language structure and composition). Consequently, language quality can be measured in terms of cognitive effects and requirements during the understanding process.

In spite of the advances in the field of neurophysiological signal acquisition (EEG, fMRI, MEG, PET, etc.) (Démonet, 2005), the measure of dynamic cognitive load and effects during language processing is still a challenge nowadays. For this reason, this paper presents a computational model of dynamic memory — Cognitive Reading Indexing Model (CRIM) — that emulates the cognitive processing of human beings during reading. Computational modeling allows monitoring and measuring the use and capacity of the internal mechanism and resources of the model. Unlike biologically inspired related models such as the Cambrias et al.'s (Cambria, Mazzocco, & Hussain, 2013), which is focused on the static extraction of emotions and polarity that a piece of text contains. The model used in this paper is based on dynamic measurements of working memory usage and capacity during essay reading. These dynamic measures are confronted with the essay scores given by human graders in order to find a correlation between the text quality and the effects on cognitive performance during reading. It is worth noting that this work is not an attempt for a better automated essay scorer. Firstly, it is a step ahead in the development of mechanisms that emulate how perceived stimuli modulate our cognitive functions (bounded cognition, Gigerenzer & Selten, 2002), which is a primary target of the BICA challenge. Secondly, it is another proof of concept on how biologically inspired models can help to give insight into the cognitive processes of the human mind.

The next section presents the most important approaches to characterize language with quantitative measures at different linguistics levels, with the aim of capturing the subjective essence of human criteria and therefore replacing human graders with automated scorers. Cognitive effects of language quality comments different psychological evidence that confirms the influence of language structure and form on the cognitive processing of comprehension, and more concretely the role that working memory (WM from now on) plays in this process. In A computational model of dynamic working memory during reading, a computational working memory model for reading is described, showing the monitoring capabilities that it offers. Materials and empirical procedure presents the experimental design and procedure to test the correlation of the essay quality and memory effects, followed by the significant results obtained. Finally, some concluding remarks and future work are discussed.

Automated measuring of language quality

One of the controversial matters regarding essay grading is subjectivity, which is thought to cause the grade variation between different human graders (Carrell, 1995). Subjectivity has often been considered as an unfair factor by students being evaluated. In order to overcome this ''problem'' as well as to save the long time spent in the essay assessment (Mason & Grove-Stephenson, 2002), automated scorers came out as a fine alternative (Valenti, Neri, & Cucchiarelli, 2003). The fundamentals of such systems is the quantification, by means of observable linguistics features, of the intrinsic variables that human raters take subjectively into account (called *trins* Hearst, 2000). For instance, the number of words of a text would represent fluency; word length variation would correlate with diction; and number of relative pronouns and different parts of speech (POS) would be related to complexity of sentence syntax (Page, 1994).

The latter mentioned features belong to the lexical and syntax levels. Other computational essay scoring systems make use of features at the semantic level. Many of them produce a statistics-based semantic representation (Leacock, 2004) of the texts and compare it with the ideal essay or master text (Jerrams-Smith, Soh, & Callear, 2001). Other systems extract features regarding the discourse/rhetorical level by measuring semantic coherence between consecutive sentences or tracking topic shifts (Burstein, Leacock, & Swartz, 2001; Higgins & Burstein, 2006; Higgins, Burstein, & Attali, 2006).

Although all these artificial scoring systems work relatively fine for concrete domains, they carry some drawbacks. Most of them apply some kind of machine learning method, which is generally supervised and therefore needs training data (Valenti et al., 2003). In this case, training data is composed of texts annotated by human subjects. Thus, training data is costly to construct, difficult to find in turn, and it is still loaded of subjectivity. In addition, most of the grading systems are optimized and evaluated against scores given by human graders. This evaluation and optimization methodology makes artificial systems overfit the concrete human graders.

The primary aim of the creation of automated essay grading systems was the ''use of computers to increase the understanding of the textual features and cognitive skills involved in the creation and comprehension of written texts'' (Valenti et al., 2003). It seems that knowledge about the correlation between textual features has been enriched since the first automated scorers. However, the same enrichment has not occurred in the cognitive counterpart.