Information Processing and Management 52 (2016) 359–373

Contents lists available at ScienceDirect

Information Processing and Management

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

Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system

Ricardo Olmos[a,c,*], Guillermo Jorge-Botana[c,b,1], José M. Luzón[c,b,1], Jesús I. Martín-Cordero[c,1], José Antonio León[a,2]

[a] Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco. C/Iván Pavlov, s/n., 28049 Madrid, Spain
[b] Universidad Nacional de Educación a Distancia, Juan del Rosal, nº 10. 28023, Spain
[c] Semantia Lab, Bravo Murillo, 38. 28015, Madrid, Spain

ARTICLE INFO

Article history:
Received 21 March 2015
Revised 26 November 2015
Accepted 1 December 2015
Available online 31 December 2015

Keywords:
LSA
Rubric
Lexical descriptors
Automatic evaluation
Summary

ABSTRACT

The purpose of this article is to validate, through two empirical studies, a new method for automatic evaluation of written texts, called Inbuilt Rubric, based on the Latent Semantic Analysis (LSA) technique, which constitutes an innovative and distinct turn with respect to LSA application so far. In the first empirical study, evidence of the validity of the method to identify and evaluate the conceptual axes of a text in a sample of 78 summaries by secondary school students is sought. Results show that the proposed method has a significantly higher degree of reliability than classic LSA methods of text evaluation, and displays very high sensitivity to identify which conceptual axes are included or not in each summary. A second study evaluates the method's capacity to interact and provide feedback about quality in a real online system on a sample of 924 discursive texts written by university students. Results show that students improved the quality of their written texts using this system, and also rated the experience very highly. The final conclusion is that this new method opens a very interesting way regarding the role of automatic assessors in the identification of presence/absence and quality of elaboration of relevant conceptual information in texts written by students with lower time costs than the usual LSA-based methods.

1. Introduction

It is a well-known fact that novice writers pay little attention to revision processes. It is also well known that even when revisions are made these are often superficial or mechanic (Fitzgerald, 1987; Graham, 2006). There is also evidence that even university students spend little time on revision processes (Pianko, 1979); in fact, it can be stated that one of the characteristics of more competent writers is spending more time on revision and applying it to a conceptual and structural level (Hayes & Flower, 1986) instead than to superficial text traits (spelling, errors, etc.) which is the common strategy in novice

* Corresponding author at: Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco. C/Iván Pavlov, s/n., 28049 Madrid, Spain. Tel.: +34 91 497 85 86.

E-mail addresses: ricardo.olmos@uam.es (R. Olmos), gdejorge@psi.uned.es (G. Jorge-Botana), jmluzon@psi.uned.es (J.M. Luzón), jimarti@psi.uned.es (J.I. Martín-Cordero), joseantonio.leon@uam.es (J.A. León).

URL: http://www.semantialab.es (R. Olmos), http://www.semantialab.es (G. Jorge-Botana), http://www.semantialab.es (J.M. Luzón)

[1] Tel.: +34 91 398 60 95.

[2] Tel.: +34 91 497 52 26.

http://dx.doi.org/10.1016/j.ipm.2015.12.002
0306-4573/

writers. The importance of improving the quality of students' writing cannot be overstated, and the same can be said about the relevance of revision processes to this purpose. Precisely in order to aid and motivate writers in the process of revision of academic texts, special emphasis has been placed in recent years on the creation of automatic assessors and tutors that provide help in instruction tasks (Foltz, Laham, & Landauer, 1999b; Foltz, Streeter, Lochbaum, & Landauer, 2013; Magliano & Graeser, 2012; Shermis, Koch, Page, Keith, & Harrington, 2002). Some of these automated assessors have made use of engines based on Latent Semantic Analysis (LSA) and its ability to identify concepts (Foltz, Gilliam, & Kendall, 2000; Graesser et al., 2000; Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007). Part of the assessors in this group are aimed at offering ongoing online conceptual feedback when a student delivers a text or changes part of a text already delivered (Franzke, Kinstch, Caccamise, Johnson, & Dooley, 2005). The aim of these assessors is to provide support to the cyclical process by which students enter the summary of a previously-read text. The system provides information about the concepts that are included or not with respect to those which should be included, and the student re-revises his text (reviewing even the basic text to obtain information). Then the student enters another summary that will usually be based on the first one.

As was previously stated, Latent Semantic Analysis (LSA) is one of the techniques that can identify conceptual information and is widely used for this reason. There is no need to specify the basics and details of this technique, which is thoroughly documented (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer & Dumais, 1997). There is extensive documentation too about the educational aspect of this technique (Haley, Thomas, Petre, & De Roeck, 2007; Kakkonen, Sutinen, & Timonen, 2005; Millis, Magliano, Wiemer-Hastings, Todaro, & McNamara, 2007). However, some essential LSA concepts should be described to understand the logic underlying the methodology proposed in this paper. LSA starts by analyzing an extensive set of documents (the linguistic corpus), where, to begin with, a matrix **X** of occurrences is generated, with n rows and m columns. The number n of rows corresponds to the number of different terms analyzed in this first step and the number m of columns refers to the number of documents used to train the LSA system. Each cell in **X** contains the frequency of the term in each document. It is known that this LSA based **X** matrix does not contain a useful representation of the semantics because, among other things, it has a huge dimensionality, and contains the subjective use made by the authors of the terms. After softening the **X** matrix by means of a function such as entropy (i.e. Nakov, Popova, & Mateev, 2001) the algebraic technique known as *singular value decomposition*, which defines the essence of LSA, is applied. This technique yields a **US** matrix with n x k dimensions, where k is now a small number of dimensions, usually ranging between 250 and 350. This **US** matrix is known as a latent semantic space because it contains an efficient representation of the terms in the corpus (i.e. Landauer & Dumais, 1997). The term "latent" is due to the fact that the semantic space has an eminently abstract nature. The k dimensions that represent terms and texts do not correspond to discernible concepts or specific episodes with which to label the dimensions. The k dimensions do not have thus psychological verisimilitude, yet they have proven very useful to simulate semantic judgments and many tasks involving text meaning (i.e. evaluation, predication, diagnosis, telephone call routing, etc.) The concept of semantic space is thus very important, as it is the basis onto which the texts to be semantically interpreted are projected, and thus is also very important for educational assessment applications.

1.1. Limitations of the current LSA based assessment systems

The way in which texts written by students have been usually assessed and rated has been to compare the vector representation of those texts in the semantic space with the vector representation of one or more texts that serve as the standard criterion and have been written by experts in the discipline (usually teachers). There are various techniques in this regard, from the simplest ones, based on a single standard summary (golden summary, see Landauer, Foltz, & Laham, 1998 or more recently Klein, Kyrilov, & Tokman, 2011), to more complex ones, such as the technique that makes use of a sample of previously rated standard summaries so that the scale is more plausible from the point of view of a human judge (grading techniques, see Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Dronen, Foltz, & Habermehl, 2015; Kakkonen & Sutinen, 2004). To the extent that the student's summary approaches a pre-rated standard summary, it will get the same grade as the standard summary. In any case, in both techniques and some intermediate ones, a single grade is obtained, which is a function of the semantic distance between the student's text and those standard criteria. Dronen et al. (2014) describe the high costs of this type of study and propose methods for pre-grading to be as efficient as possible.

One of the ways in which these techniques have been implemented to assess student texts has implied providing not only a single grade, but the information of what contents are missing in a summary to reach an acceptable standard as well. This is basically a rubric-based content detection task (see, for instance, *Summary Street* by Franzke et al., 2005). This is done in the same way as when giving a single grade, that is, by taking text samples (or partial golden summaries) from the concepts to inform about (e.g. sentences extracted from the book to be read), project them onto the semantic space as a vector, and compare them to the vector that represents the student's text. The feedback for each concept will depend on whether the distance between vectors (those of the students and those of the partial golden summaries) is large or small. Another example is Apex (Dessus & Lemaire, 1999), an interactive learning environment, in which the teacher must identify short passages of the instructional text that contains a key concept and the topic or topics to which each concept belongs. In Magliano and Graeser (2012) it is shown that automatic systems involve the creation of a number of answers (which they call expectations) that represent either the different levels of students, or else concepts, inferences, or usual misconceptions among students. These answers are real language samples. Obviously, the preparation of these instruments (golden summaries, graded summaries or partial golden