Feature extraction and LSA in automatic essay evaluation

João Santos[1], Eloi Favero[2], Silvério Sirotheau[3], and Simone Freitas[4]

[1]*Faculty of Mathematics – Federal University of Pará – Belém - Pará – Brasil,* jcas@ufpa.br
[2]*Faculty Computer Science – Federal University of Pará – Belém - Pará – Brasil,* favero@ufpa.br
[3]*Faculty Petroleum Engineering – Federal University of Pará – Salinópolis - Pará – Brasil,* ssirotheau@gmail.com
[4]*Faculty of letters – Federal University of Pará – Castanhal - Pará – Brazil,* sinegrabr@gmail.com

1 Abstract

The technological advances in the field of education have enabled a variety of online activities, among which are the answers to discursive questions. Such questions serve the purpose of verifying and gauging students' learning, in particular the abilities to write and structure the argumentative discourse. Within this context, the development of approaches to foster automate correction gains relevance, as it presents advantages such as:

i) it reduces the correction time, proving more stable in the correction procedure, reducing the possibility of errors;
ii) it provides immediate feedback which can support students in solving questions;
iii) it allows multiple submissions at a time and place determined by the user;
iv) it has low cost.

The main focus of the experiments here presented is the development of a set of algorithms for the automatic evaluation of written answers to discursive questions based on Natural Language Processing and Artificial Intelligence techniques. The main techniques used:

1. The initial work was to apply a latent semantic analysis (LSA) model for the automatic evaluation of short answers (25 to 70 words) to open-ended questions of the entrance examinations from the Federal University of Pará (Ufpa);

2. Next, we conducted experiments using a method based on n-gram similarity and a categorization process;

3. Next, we conducted experiments using a method based on machine learning techniques to a corpus constituted of a sample covering 1000 essays from a public exam for admission to the career of Administrative Technician in Education.

A system is considered to perform well when the system versus human agreement index approaches or is higher than the human versus human agreement index. Considering a study corpus constituted of 1,000 (thousand) essays obtained from a selection process and previously evaluated by two human specialists, the quadratic kappa index (QK) among human evaluators was 0.68. For this same corpus, in our experiments, we obtained a system versus human index QK of 0.8412. In light of the results, we can consider that this technology is achieving a good performance, thus, the method can be used in combination with human evaluation in real tests with written answers to discursive questions.