Automatic evaluation of discursive answers short based on three linguistic dimensions

Avaliação automática de respostas discursivas curtas com base em três dimensões linguísticas

DOI:10.34117/bjdv7n8-474

Recebimento dos originais: 07/07/2021
Aceitação para publicação: 19/08/2021

Silvério Sirotheau Corrêa Neto
Doutor
Campus Universitário de Salinípolis - UFPa
Rua Raimundo Santana Cruz, S/N - São Tomé - Salinópolis - PA
silverio@ufpa.br

Eloi Luiz Favero
Doutor
Instituto de Ciências Exatas e Naturais – UFPa
Rua Augusto Corrêa, 01 - Guamá. - Belém - PA
favero@ufpa.br

João Carlos Alves dos Santos
Doutor
Instituto de Ciências Exatas e Naturais – UFPa
Rua Augusto Corrêa, 01 - Guamá. - Belém - PA
jcas@ufpa.br

Marco Aurélio Lima do Nascimento Júnior
Graduando
Instituto de Ciências Exatas e Naturais – UFPa
Rua Augusto Corrêa, 01 - Guamá. - Belém - PA
marconasc505@gmail.com

Simone Negrão de Freitas
Mestre
Campus Universitário de Castanhal
Av. dos Universitários - Jaderlândia - Castanhal - PA
negrao@ufpa.br

ABSTRACT

As the use of virtual environments grows, there is a need for a system of automatic evaluation of discursive answers. This paper proposes a method for automatic evaluation of discursive short answers based on a machine learning architecture. The predictive method is based on the collection of features (140) of similarity between texts in a taxonomy of three linguistic dimensions: lexical, syntactic and semantic. As a result, we obtained quadratic kappa 0.72 human x system (SxH) against 0.94 human x human (HxH)

for the proof of biology and an accuracy 0.76 SxH against 0.58 HxH for the proof of geography.

Keywords: automatic evaluation, machine learning, linguistic dimensions, features

RESUMO
À medida que o uso de ambientes virtuais cresce, surge a necessidade de um sistema de avaliação automática das respostas discursivas. Este artigo propõe um método de avaliação automática de respostas discursivas curtas baseado em uma arquitetura de aprendizado de máquina. O método preditivo é baseado na coleção de características (140) de semelhança entre textos em uma taxonomia de três dimensões linguísticas: lexical, sintática e semântica. Como resultado, obtivemos kappa quadrático 0,72 humano x sistema (SxH) contra 0,94 humano x humano (HxH) para a prova de biologia e uma precisão de 0,76 SxH contra 0,58 HxH para a prova de geografia.

Palavras-chave: avaliação automática, aprendizado de máquina, dimensões linguísticas, recursos

1 INTRODUCTION

During their school path, the student undergoes a continuous evaluation process, cumulative and systematic. Even in the face of more modern pedagogical conceptions, evaluation applications composed of discursive questions have strong relevance, as they assess student learning outcomes, in particular their writing skills and understanding of specific concepts in a given domain (Zupanc and Bosnic, 2017; Page, 1966).

However, the manual correction task of this evaluation type for a large number of students is very costly in terms of human resources, time and money. For example, the National Secondary Education Examination (ENEM) which is a selective process to enter federal institutions of higher education in Brazil, with more than 6 million candidates, has in its structure essay-argumentative text questions. What is the amount of time and cost to evaluate more than 6 million texts?

Rababah and Al-Taani (2017) state that manual correction can consume a lot of teacher's time and what computer systems can help in this type of task. This type of system contributes to helping the human evaluator, releasing him in part from the manual correction, so he can direct his attention to more specific points of the teaching-learning process. In this context, the development of algorithms for automating the correction of answers to discursive questions becomes very relevant in the teaching-learning process (Pérez et al. 2005).

In the field of automatic evaluation of short discursive answers there are two main lines of research: (1) The first is based on corpus and similarity between texts (Gomaa and Fahmy 2014; Pribadi et al., 2017) and; (2) The second is based on metrics of

similarity between networks of concepts extracted from the texts of the answers using machine learning techniques and natural language processing (PLN) (Mohler and Mihalcea 2009; Zupanc and Bosnic, 2017; Palma and Atkinson, 2018).

In the approach based on text similarity the PLN is only superficial (token collection) while in the similarity approach between concept networks more sophisticated PLN and Machine Learning methods are needed (labelling, pronoun resolution, entity extraction, among others).

This work proposes a method for automatic evaluation of short answers based on a 5-step machine learning pipeline. The predictive method is based on the collection of features (140) of similarity between texts in a taxonomy of three linguistic dimensions: lexicon, syntactic and semantic. One of our contributions is working with these features, originally proposed for other languages, directing the research to the Portuguese language. The aim is to achieve an accuracy value close to that obtained between two human evaluators (HxH). When a system contrasted with humans reaches accuracy values close to those of human evaluators (HxH), it becomes reliable to be used in correcting discursive answers (Haley et al. 2007).

This research contributes to generate innovative technology for automatic evaluation of short discursive answers. This technology in virtual learning environments has the advantages: (i) immediate feedback for the student, even in a very large number of students; (ii) low financial cost; allows multiple evaluations in one interactive response development; (iii) uniformity in the evaluation, as it is independent the evaluator's physical and emotional fatigue; (iv) releases the teacher from manual correction, allows it to direct more attention to specific points.

This article is organized as follows: Section 2 presents related works. Section 3 presents the methodology. Section 4 presents results and discussion and section 5 presents the conclusion.

2 RELATED WORKS

Research on the automatic evaluation of texts (long answers) started in the 1960s with the PEG system, with a focus on assessing writing style skills of students (Page 1966). Later other initiatives emerged from the 90s, with the emergence of PLN techniques providing a considerable advance in these fields such as E-rater (Burstein et al. 1998) and Intellimetric (Learning 2000).

More recent efforts have been achieving an accuracy very close to the measure among human evaluators. Leacock and Chodorow (2003) describe a mechanism for score of short discursive answers from a reference answer made by experts, combining syntactic features of a student response (subject, object and verb) with a set of reference responses. They worked with a corpus of 16,625 answers, achieving an agreement accuracy of 84% against the humans evaluators.

Mohler and Mihalcea (2009) explored unsupervised techniques of machine learning for automatic assessment of short answers. They combined knowledge-based measures from WordNet and Latent Semantic Analysis (LSA). They achieved a correlation of 0.50 (S×H) against a correlation of 0.64 of (H×H).

Gomaa and Fahmy (2014) used several similarity metrics (String-based similarity, Corpus-based similarity, Knowledge-based similarity, Hybrid similarity measures and Sentence-level semantic similarity) as input to the classification method; on a corpus of 610 short answers from students rated on a scale of 0 to 5, and obtained a correlation of 0.68 (S×H) against a correlation of 0.86 (H×H).

Rodrigues and Araújo (2012) explored PLN techniques with a step of translation of phrases to canonical forms (word lists vs. tag) via substitution of synonyms, such as the use of a thesaurus. In the classification step, they used the vector space model and achieved a correlation of 0.78 between the evaluators' mean and the score given by the system.

Galhardi et al., (2018) present a new database for short answers in Portuguese, with an approach composed of 4 groups of features (bag of n-grams, lexical similarity, semantic similarity and statistics) obtaining the results using Extreme Gradient Boosting Classifier and cross-validation, achieving an accuracy of 69% and a kappa agreement of 0.54.

3 RESEARCH QUESTIONS

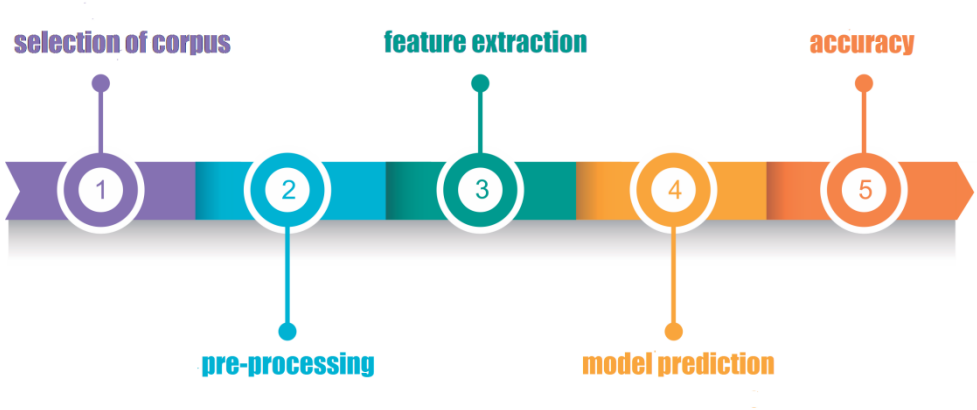
In the bibliographical survey on short-type discursive answers, some questions (Q): I - Among the various pre-processing techniques (Burrows et al. 2015), in this work three are used: surface (ex. punctuation removal), lexicon (eg spelling correction and stop word removal), morphological (eg stemmer). (Q1) the pre-processing does influence the final accuracy in this type of approach? II - Vajalla (2018) states that little is known about which linguistic features are good predictors. (Q2) which are the best predictive features

for the Portuguese language in matters of short discursive answers? (Q3) The contribution importance of the features is repeated on different datasets?

4 METODOLOGY

The approach is centered on a pipeline architecture that contains 5 steps: (1) selection of corpus, (2) pre-processing, (3) feature extraction, (4) model prediction and (5) accuracy (see Figure 1).

Figure 1. Pipeline architecture for evaluating short texts.



In the corpus selection step, we selected two datasets with short answers for Portuguese, related to two questions arising from a college entrance test. One question is Biology with 130 answers and the other is Geography with 229 responses.

In the pre-processing stage, the responses were vectorized into sentences and then separated into tokens. After that, three pre-processing techniques were used: (1) Removal of Special Characters, punctuation, accent and conversion of uppercase letters to lowercase letters (RCE); (2) Removal of stop word (RSW) and; (3) Removal of suffixes (stemmer) (RSU). For pre-processing we use the library Natural Language Toolkit (NLTK), where the techniques were combined in the following form: a) without preprocessing (-RCE, -RSW, -RSU); b) with removal of special characters (+RCE, -RSW, -RSU); c) with removal of special characters and stop words (+RCE, +RSW, -RSU) and; d) with removal of special characters, stop words and the application of stemmer (+RCE, +RSW, +RSU). Then the tokens were tagged morphologically for classification according to their grammatical categories, for this we use Aelius (Alencar, 2010).

In the feature extraction step (attributes, characteristics or text variables), we tried to cover all the main attributes in recent literature (Zupanc and Bosnic, 2017; Palma and

Atkinson, 2018; Vajjala, 2018) (see table 1). 140 features were extracted and grouped in 3 dimensions: lexical, syntactic and semantic.

Lexical Dimension. It collects features that describe the individual aspect of words. In this dimension we have 4 main categories: (1) surface statistics, collects statistics based on word count. (2) diversity, collects measures that represent how diverse the vocabulary used is. (3) readability, measures the degree of ease of reading the text. (4) Error, number of spelling errors.

Syntactic dimension. It collects features that portray the individual aspect of each sentence, comprises two categories: (1) number of each PoS tag (part-of-speech tagging), such as number of nouns and verbs (2) Lexicon and Syntactic error, counts the number of errors in poorly worded sentences, for example, errors in agreement and punctuation.

Semantic Dimension. It collects features that describe the aspects that are related to text content, for example, similarity measures between the student's and the reference answer. And, it also collects features that describe the aspects related to textual coherence, both local within a response and global in relation to the various answers.

TABLE 1. Taxonomy of features for automatic evaluation of texts in the language Portuguese, for short answers (part 1/2)

Lexical	Statistics of Surface	n° of characters, n° of different words, n° of words, n° of words short, n° of long words, n° of average words, n° of stop word, n° sentence, n° most frequent word length.
	Diversity	Type-token-ratio – TTR, Guiraud's index, Yule's K, the D estimate, hapax legomen.
	Redability	Gunning Fox Index, Flesch Kincaid grade level, Dale-Chall readability formula, autometed readability index, LIX, word variation index, nominal ratio, SMOG-index
	Error	number of spelling errors
Syntactic	n° of each PoS tags	Number of different PoS tags, Number of tags per syntactic category:SR=being, HV=haver, ET=being, TR=having, VB=verb (-I=imperative, -P=present, -SP=present subjunctive, -D=past, -RA=inflectional phoneme, -SD=past subjunctive, -R=future, -SR=future subjunctive, -G=gerund, -PP=perfect participle, -NA=agrement particle), Agreement Particle (genre(none=masc,-F=fem, -G=double gender), number(none=sing, -P=plural)) , N (noun) NPR (proper noun) PRO (pronouns) P+PRO(Prep+Pronoun) PRO$ (possessive) CL (clitics) D (determine) DEM(demonstrative) ADJ (adjective) ADV (adverbs) Q (quantifier) CONJ(conjunction) C (subordinating conjunction) WPRO (relative) WQUE(interrogative) WD (interrogative determiners) P(preposition)
	Error	n° of punctuation errors

In collecting content features a student answer is contrasted with the reference answer, common in the use of n-grams (uni and bi) with distance measurements, Euclidean and cosine. We also use local and global weighting methods for texts like tf-idf. Typically the reference answer is formed from a set of the most highly rated answers.

On the other hand, some authors suggest that one can also use responses from reference based on groupings made in relation to the score (Zupanc and Bosnic, 2017). Based on this, considering the scores in the range 0 to 6, we created 7 vectors, one reference answer for each score value. Here we apply the measurements (distance Euclidean and cosine) against these response vectors, also including variations in the type of pre-processing; this resulted in 66 content features, many of which are in the most relevant ones. Still in the semantic dimension, we assess the coherence of the text, which describes the flow of information within text. For this, we use an approach based on overlapping windows (Zupanc and Bosnic, 2017; Palma and Atkinson, 2018). We used 4 models that generated 66 features, as shown in table 2.

TABLE 2. Taxonomy of features (part 2 / 2).

Semantic**	Content	Similarity* Cosine and distance Euclidean with reference answer	Similarity with Source Text (Pre: SSW, CST, CSW) (Med: Cosine and Euclidean Distance)
		Similarity and distance against the score ranges	Similarity (level: 0, 1, 2, 3, 4, 5, 6)(Pre: SSW, CST, CSW) (Med: Cosine and Euclidean Distance)
		Weighted sum of all correlation values based on values of Cosine and Euclidean distance	Weighted Correlation (Pre: SSW, CST, CSW) (Med: Cosine and Euclidean Distance)
	Coherence	Distances between two adjoining windows	Min, med, max
		Distances from all windows against all	
		Local center, all windows against the local center	
		Global center, all windows against the global center	

Similarity* Cosine and Euclidean distance: actually cosine is a measure of similarity, while the Euclidean distance is a measure of dissimilarity. To make the two measurements like similarity we consider 1/Euclidean distance.
** the number of features is given by the simple multiplication of each item group, for example, for content we have 7x3x2=42

In the prediction step, we use the Random Forest algorithm, which as a supervised machine learning method allows the combination of hundreds of features in regression and/or prediction tasks. It creates a set of decision trees, where each tree is trained by a

different subset of data from the training set. For this type of problem, where we have a large number of features, more than 100, the Random Forest algorithm performs well (Fernández-Delgado et al. 2014). For validation we use the Cross-validation approach, partitioning the set of data in 5 folds; the accuracy collected is the average of the 5 tests.

In the accuracy step, we seek to select the best combinations of the prio steps to maximize accuracy. To measure accuracy we used Kappa Quadratic - KQ (Fleiss and Cohen, 1973), which measures the degree of agreement between two classes with a certain flexibility regarding exact agreement. KQ measures also the partial agreement: if it should predict 6, but it resulted in 5, it is not totally wrong. This metric usually ranges from 0 (little agreement between evaluators) to 1 (complete agreement between evaluators). If the agreement between the evaluators is below the expected minimum, this metric can also result in negative values.

KQ is calculated by creating a matrix according to equations 1 and 2. In this case, the matrix O contains the scores, such that the classification i is given by the human evaluator and j given by the model. $W_{i,j}$ contains the weights as derived in the Equation 1 and the matrix E contains the scores expected from the human evaluators, obtained by the multiplication of the histogram vectors of the two scores. Subscripts in matrix $O_{i,j}$ correspond to the number of answers that score i from the human evaluator and j from the system.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

At the end of the KQ process it is calculated as:

$$k = 1 - \frac{\sum_{i,j} W_{i,j}\, O_{i,j}}{\sum_{i,j} W_{i,j}\ E_{i,j}}$$

The interpretation of KQ results, between 0 and 1, between little and a lot of agreement, can be somewhat subjective. So we quote an interpretation recommended by Landis and Koch (Landis and Koch, 1977) which considers six tracks of values: i) < 0.00 → “Poor”, ii) 0.00 - 0.20 → “Weak”, iii) 0.21 - 0.40 → “Reasonable”, iv) 0.41 - 0.60 → "Moderate", v) 0.61 - 0.80 → "Substantial" and vi) 0.81 - 1.00 → "Almost Perfect".

5 RESULTS AND DISCUSSION

Our research corpus consisted of a collection of answers to two discursive questions contained in the public notice 016/2007 of the entrance exam from the Federal University of Pará. From a universe of one thousand answer sheets, the two questions with more completed answer sheets: Biology with 130 answers and Geography with 229 answers. The candidate chose a subset of the 26 questions he would answer, which explains why we don't have a thousand answers for every question. The Biology question has averages 28 words per answer and the Geography 74 words. Each answer has a score of two human evaluators, so we can calculate the accuracy of hits between them (H×H).

The approach was applied with the goal of maximizing the S×H value looking for an approximation with H×H. Table 3 presents the results in KQ.

Table 3. Results of Biology and Geography short answers.

Data base	**SxH**	**HxH**
Biology	0.72	0.94
Geography	0.76	0.58

This table consolidates the results that are promising. For Biology the HxH was of 0.94, in the above KQ interpretation is an almost perfect agreement. The system reached an SxH value of 0.72, which is substantial agreement. In the Geography question the HxH was 0.58 which is a moderate agreement, however the system achieved an SxH 0.76, which is a substantial agreement, outperforming the HxH.

In relation to the research question raised (Q1), where Burrows et al., (2015) report various pre-processing techniques used in word processing. Three morphological processing techniques were used: (1) Removal of Special Characters and Punctuations (+RCE); (2) removal of stop words (+RSW); and (3) removal of suffixes (stemmer) (+RSU). These three techniques were combined in four ways: i) without pre-processing (-RCE, -RSW, -RSU), with removal of special characters (+RCE, -RSW, -RSU), with removal of special characters and stop words (+RCE, +RSW, -RSU) and with removal of special characters, stop words and stemmer application (+RCE, +RSW, +RSU). In table 4 we have the results obtained for Biology and Geography considering variations in pre-processing techniques.

Table 4. Pre-processing of short answers from Biology and Geography.

	Biology		Geography	
human vs. human	0.94		0.58	
Average words per answer	28.48		74.56	
System vs. Human	Cont	Lex+Sint+Cont	Cont	Lex+Sint +Cont
-RCE, -RSW, -RSU	0.65	0.64	0.70	0.70
+RCE, -RSW, -RSU	0.70	0.70	0.74	0.71
+RCE, +RSW, -RSU	0.64	0.64	0.66	0.76
+RCE, +RSW, +RSU	0.71	0.72	0.73	0.69

As shown in Table 4, the different pre-processing techniques present different accuracy values. However, the differences are quite significant within each base, with the difference from the smallest to the largest value being 0.08 in Biology and 0.10 for Geography, which answers research question Q1. Considering these values is It is important to have the pre-processing step in the automatic evaluation approaches of short answers.

In the discussion of research question Q2, on what are the best feature predictors for the Portuguese language in questions of short discursive answers? In table 5 we present the main features in order of importance. We started with a set of more than 140 features, and we used the random forest method for prediction and selection of the importance of features.

Table 5. Importance of features in Portuguese short answers.

Features	Importance	Features	Importance
Cosseno Escore 4	0.23	Distância Euclidiana Escore 0	0.06
Cosseno Escore 6 sem Stop Word (SW)	0.13	Cosseno Escore 6	0.05
Cosseno Escore 5	0.12	Cosseno Escore 3	0.05
Número de caracteres	0.09	Cosseno Escore 3 com SW	0.05
Cosseno Escore 5 sem SW	0.09	Número de palavras	0.05
Cosseno com texto fonte e com SW	0.07	Cosseno Escore 4 com SW	0.04
Número de stop word	0.06	Número de pronomes	0.04
Número de palavras longas	0.06	Número diferente de palavras	0.03

Regarding the research question (Q3) The importance of the contribution of features repeats itself in the different data sets? In table 6 we selected the best features of each base and we verify that the cosine and Euclidean distance measures per score range are the main features of the two bases. On the other hand, the other features most relevant are surface statistics lexicons such as number of words. and number of different words and number of long words.

Table 6. Result of the importance of features in each database.

	Biology		Geography	
N°	**Features**	**Importance**	**Features**	**Importance**
1	cosseno escore 6 sem stop word	0.13	cosseno escore 4	0.23
2	cosseno escore 5	0.11	euclidiana escore 0	0.06
3	cosseno escore 5 sem stop word	0.09	cosseno escore 3	0.05
4	número de caracteres5	0.09	número de stop word	0.05
5	cosseno com texto fonte	0.07	cosseno escore 3 com stop word	0.04
6	cosseno escore 6	0.05	cosseno escore 4 com stop word	0.04
7	número de palavras longas	0.04	número de palavras	0.03
8	cosseno texto fonte sem stop word	0.03	número de palavras diferentes	0.03
9	número de pronomes	0.03	cosseno escore 2	0.02
10	cosseno escore 4 sem stop word	0.02	cosseno escore 2 com stop word	0.02

6 CONCLUSION

The objective of this work is to develop an automatic evaluation method for short discursive answers based on the similarity between texts, collecting features in three main dimensions: lexicon, syntactic and semantic. They have been classified in a kind of taxonomy with over 140 features. Most of them came from related works from the English language which have been adjusted to Portuguese. To carry out the experiments a 5-step linear pipeline architecture was used: corpus selection, pre-processing, feature extraction, prediction model and accuracy.

From the values of the features collected, the objective is to predict the value of the score of each answer with an accuracy close to that measured between two human evaluators. We use the Random Forest technique, which allows the manipulation of a large number of features in addition to returning the relevance of each feature in the classification step. As a result we obtained a quadratic kappa (KQ) 0.72 SxH against 0.94 HxH for the Biology test and an SxH value of 0.76 against HxH 0.58 for the Geography test. A KQ score of 0.72 is considered “substantial” even though it is lower to the one collected between two human evaluators. On the other hand, in the Geography test the system with 0.76, also "substantial", surpasses the accuracy measured between the two human evaluators that was 0.58, a "moderate" value. This result shows that this technology is reaching a state of maturity to be used in virtual learning environments.

REFERENCES

Alencar, L. F. (2010) "Aelius: uma ferramenta para anotação automática de corpora usando o NLTK", Anais do IX Encontro de Linguística de Corpus, PUCRS, Porto Alegre, v. 8.

Burrows, S., Gurevych, I. and Stein, B. (2015) "The eras and trends of automatic short answer grading", International Journal of Artificial Intelligence in Education, v. 25, n. 1, p. 60-117.

Burstein, J. et al. (1998) "Automated scoring using a hybrid feature identification technique", In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pages 206–210. Association for Computational Linguistics.

Fernández-Delgado, M. et al. (2014) "Do we need hundreds of classifiers to solve real world classification problems?". The Journal of Machine Learning Research, v. 15, n. 1, p. 3133-3181.

Fleiss, J. L. and Cohen, J. (1973) "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability", Educational and psychological measurement, v. 33, n. 3, p. 613-619.

Galhardi, L. et al. (2018) "Portuguese Automatic Short Answer Grading". In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). p. 1373. 1559 Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019) VIII Congresso Brasileiro de Informática na Educação (CBIE 2019)

Gomaa, W. H. and Fahmy, A. A. (2014). "Automatic scoring for answers to arabic test questions". Computer Speech & Language, 28(4):833–857.

Haley, D. T. et al. (2007) "Seeing the whole picture: evaluating automated assessment systems". Innovation in Teaching and Learning in Information and Computer Sciences, v. 6, n. 4, p. 203-224.

Landis, J. R. and Koch, G. (1977) "The measurement of observer agreement for categorical data". biometrics, p. 159-174.

Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 37(4):389–405.

Learning, V. (2000). "A study of expert scoring and intellimetric scoring accuracy for dimensional scoring of grade 11 student writing responses" (rb-397). Newtown, PA: Vantage Learning.

Mohler, M. and Mihalcea, R. (2009). "Text-to-text semantic similarity for automatic short answer grading". In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 567–575. Association for Computational Linguistics.

Page, E. B. (1966). “The imminence of... grading essays by computer”. The Phi Delta Kappan, v. 47, n. 5, p. 238-243.

Palma, D. and Atkinson, J. (2018) “Coherence-Based Automatic Essay Assessment”. IEEE Intelligent Systems, v. 33, n. 5, p. 26-36.

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C., and Magnini, B. (2005). “About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment”. Revista signos, 38(59):325–343

Pribadi, F. S., Adji, T. B., Permanasari, A. E., Mulwinda, A., and Utomo, A. B. (2017). “Automatic short answer scoring using words overlapping methods”. In AIP Conference Proceedings, volume 1818, page 020042. AIP Publishing.

Rababah, H. e Al-Taani, A. T. (2017) “An automated scoring approach for Arabic short answers essay questions”. In: 8th International Conference on Information Technology (ICIT). IEEE, p. 697-702.

Rodrigues, F. and Araújo, L. (2012) “Automatic Assessment of Short Free Text Answers”. In: CSEDU (2). p. 50-57.

Vajjalla, S. (2018) “Automated assessment of non-native learner essays: Investigating the role of linguistic features”. International Journal of Artificial Intelligence in Education, v. 28, n. 1, p. 79-105.

Zupanc, K. and Bosnic, Z. (2017) “Automated essay evaluation with semantic analysis”. Knowledge-Based Systems, v. 120, p. 118-132.