Journal of Computer Science

Original Research Paper

Automated Evaluation of Short Answers Using Text Similarity for the Portuguese Language

[1]Silvério Sirotheau, [2]Joao C.A. dos Santos, [1]Eloi L. Favero and [3]Simone N. de Freitas

[1]Postgraduate Program in Computer Science, Federal University of Pará, Belém - PA, Brazil
[2]Faculty of Mathematics, Federal University of Pará, Belém - PA, Brazil
[3]Faculty of languages, Federal University of Pará, Castanhal - PA, Brazil

Article history
Received: 07-06-2019
Revised: 16-08-2019
Accepted: 20-11-2019

Corresponding Author:
Silverio Sirotheau
Postgraduate Program in Computer Science, Federal University of Para,
Belem-PA, Brazil
Email: ssirotheau@gmail.com

Abstract: Along with the development of distance education, emerges the demand for virtual environments as the automated evaluation studies of essays that has already produced promising results. However, when dealing with short answers, replicating the decisions of a human grader is still a challenge, as the portability of essay evaluation techniques to short answers has not produced results with the same level of accuracy. In this sense, the present paper aims to foster the development of studies in the field of automated evaluation of short discursive answers. The related works presented three main approaches: text-to-text similarity, knowledge-based similarity that rely on synonym dictionary and corpus-based similarity that rely on a related corpus. The present study has employed an n-gram based similarity and a categorization process applied to three sets of answers to questions in Portuguese language: two of them (Biology and Geography) obtained from an admission process to higher education and the third (Philosophy) from a virtual learning environment. The employed method was comprised of a five-stage pipeline architecture: corpus selection, preprocessing, variable generation, classification and accuracy validation. In these three corpora, several similarity measurements and distances resulting from the unigrams/bigrams combination were explored. During the classification stage, two methods were used: multiple linear regression and K-Nearest Neighbors (KNN). At the same time some research questions were revised leading to meaningful findings. As for the system efficiency regarding the Biology corpus, the accuracy was 84.01 system vs. human compared to 93.85 human vs. human; for the Geography corpus, the accuracy was 86.29 system vs. human compared to 84.93 human vs. human; and for the Philosophy corpus, findings revealed 81.59 accuracy system vs. human. These results, when compared with those obtained from recent experiments produced by other techniques indicate advantages in terms of a simpler method added to good accuracy.

Keywords: Automated Evaluation, Discursive Answers, N-Grams, KNN, Linear Regression

Introduction

Evaluations of discursive answers are of great relevance, as they assess learning outcomes, emphasizing students' performance in writing, including higher-order thinking skills, such as synthesis and analysis (Magnini *et al*., 2005; Zupanc and Bosnic, 2017; Shermis *et al*., 2002). Considering these aspects, automated evaluation may represent an essential tool in learning environments.

It is possible to anticipate the resulting impact of applying one of these systems in terms of saving time and money when used for assessing essays in an admission process with over 7 million candidates, as in Brazil's ENEM-Exame Nacional do Ensino Médio (INEP, 2018). Unlike humans who naturally go through fatigue in face of demanding, arduous tasks such as the correction of many discursive answers, machines do not undergo effectiveness reduction under the same circumstances. Advantages of automated evaluation include: (i) time and cost savings: (ii) instantaneous

© 2019 Silvério Sirotheau, Joao C.A. dos Santos, Eloi L. Favero and Simone N. de Freitas. This open access article is distributed under a Creative Commons Attribution (CC-BY) 3.0 license.

feedback, allowing students multiple submissions; (iii) assisting large-numbered groups of students in virtual classrooms and (iv) supporting professor's assessment task, thus enabling him to focus on other more specific aspects on account of reducing manual correction.

Researchs on written texts automated evaluation has been underway since the sixties (Page, 1966; Hearst, 2000; Noorbehbahani and Kardan, 2011), producing a variety of systems, especially for scoring essays (long discursive answers), as we can see in the examples bellow:

E-rater (Burstein *et al*., 1998) relies on statistical surface feature models (word frequency/sentences, grammar mistakes, readability etc.) as well as on Natural Language Processing (NLP) techniques; its adjacent agreement achieves values ranging from 0.87 to 0.94 (difference of 1 point in a six-point scale).

Intelligent Essay Assessor (IEA) (Foltz *et al*., 2013) uses Latent Semantic Analysis (LSA) and also statistical surface features (over 60 attributes); with 0.70 average correlation (essay number = 1.205).

Intellimetric (Rudner *et al*., 2006) uses statistical surface features and a combination of Artificial Intelligence (AI) and Natural Language Processing (NLP); it gathers over 400 semantic, syntactic and discursive attributes; 0.86 correlation coefficient (essay number = 493).

Semantic Automated Grader for Essays (SAGE) (Zupanc and Bosnic, 2017) combines statistical surface features, content analysis and semantic coherence. It has over 80 attributes and 0.9 accuracy (kappa quadratic weighing).

These systems achieved promising results and produced accuracy levels that surpass human vs. human accuracy (Attali *et al*., 2010; Zupanc and Bosnic, 2017). On the other hand, automated evaluation of short answers, despite using the same techniques employed at essay scoring, has not achieved satisfactory performances (Magnini *et al*., 2005; Pribadi *et al*., 2017) and replicating decisions of a human grader for short answers remained a challenge.

In the related work on the automatic assessment of short answers, we find three main approaches: text-to-text similarity; knowledge-based similarity relying on a synonym dictionary (e.g., WordNet) to expand the vocabulary; and corpus-based similarity relying on a related corpus to expand the vocabulary.

This study will employ corpus-based similarity. Therefore, it may be applied to any domain and is more flexible to accommodate improvements of assessment procedures based on Machine Learning techniques, among which the major ones are the Vector Space Model (Salton *et al*., 1975) and the Latent Semantic Analysis (Landauer *et al*., 1998).

It is held that as the system accuracy level approaches that of human graders, the former becomes reliable for practical use (Trusso Haley *et al*., 2007). Thus, this research sought to develop an assessment method to automatically score short discursive answers, which presented system vs. human accuracy close to the accuracy verified between two graders (human vs. human).

Research outcomes point to an 84.01 accuracy level in system vs. human against 93.85 in human vs. human for answers to Biology questions. For answers to Geography questions, the accuracy was 86.29 system vs. human, against 84.93 accuracy in human vs. human; for answers to the Philosophy questions, 81.59 accuracy in system vs. human was verified (in this last case, human scoring was performed by a single grader). Besides the outcomes themselves, in the following sections of this paper we show how certain experiments, taking place during the elaboration of the method, led to some findings – for example, concerning the nature of the reference answer or the potential for portability of the method across linguistic domains. Moreover, they also raised issues such as the existing relationship between the distinct types of questions (conceptual or argumentative) and the features of the method.

The overall organization of the paper is as follows: first section, introduction; second section, related works; third section, defines the research corpus; fourth section, approaches the employed method; fifth section, results and discussion; and sixth section, presents our conclusions.

Related Works

Despite advances in automatic essays evaluation research, dealing with short answers is still a challenge to replicate the accuracy of human scoring. Here we report some of the related works.

Pribadi *et al*. (2016) used a simple word overlap similarity method, based on the Sorensen-Dice coefficient, the Jaccard index and cosine coefficient. They conclude that "similarity measurement cannot rely only on word overlap", because short answers have a limited numbers of words.

Mohler and Mihalcea (2009) worked with 630 answers from three assignment with 7 questions. The students' answers were collected from an online learning environment, from a computer science course at the University of North Texas (Texas corpus). The answers were independently graded by two human judges, using an integer scale from 0 to 5. The correlation between the two humans judges, per-assignment setting, was r = 0.72. In their experiment, they achieved a result r = 0.67 per-assignment setting. This result was obtained by a corpus-based measure, combining text similarity techniques over an extended corpus, which included the Computer Science subset from Wikipedia.

Benomran and Ab Aziz (2013) run another experiment using a portion of the Texas corpus, 360 answers, 3 assignments, 120 answers/assignment (a subset of the 630 answers of the above-mentioned experiment).

Table 1: Characteristics of research corpora

Corpus	Quantity	Words	H × H
Biology	131	Min = 4, Max = 56, Average = 28.48	93.84
Philosophy	192	Min = 11, Max = 269, Average = 149.37	-
Geography	230	Min = 9, Max = 189, Average = 74.56	84.93

They used a two-step approach: the vocabulary is expanded with a synonym dictionary (knowledge-based similarity), then, three complementary metrics (subsequent commons words, longest common and distance semantics) are combined by linear regression. The per-assignment correlation between the two human evaluators was (*r* = 0.72). Their approach achieved a correlation coefficient of 0.82 with the human grader.

Santos and Favero (2015) proposed an application of a Latent Semantic Analysis (LSA) model to automatic evaluation of short answers (25 to 70 words) to open-ended questions. The procedures consisted in the following tasks: first, the implementation of words bigrams; second, the implementation of combined models of unigrams and bigrams using multiple linear regression; and finally the addition of an adjustment step after the score attribution, considering the average of words per answers. Among the results, the experiments produced accuracy about 84.94, while the accuracy between two human evaluators was about 84.93.

In summary, were found three main approaches for assessing short answers: the first one is based on corpus and text similarities (Gomaa and Fahmy, 2012; Santos and Favero, 2015; Pribadi *et al*., 2017); the second one is knowledge-based similarity that rely on synonym dictionary (Benomran and Ab Aziz, 2013); the third one is corpus-based similarity that rely on a related corpus to expand the vocabulary (Mohler and Mihalcea, 2009).

There is a fourth one approach based on concept-similarity extracted from answers, using machine learning procedures and NLP such as tagging and extraction of triples – SVO (Mitchel *et al*., 2002; Sukkarieh *et al*., 2003; Mohler and Mihalcea, 2009).

The present study will employ an n-gram based similarity and a categorization process (multiple linear regression and k-nearest neighbors - KNN) applied to three sets of answers to questions in Portuguese language. This approach can be applied to another domain.

Corpus

The research corpora comprehends three data sets in Portuguese language classified as short answers obtained from the following sources: two questions were extracted from an admission process for higher education, one of which, on Biology, requiring the elaboration of three concepts within a wider scope of cytological taxonomy and the other question, on Geography, requiring the elaboration of an argument advocating a certain point of view concerning to Human and Economic Geography in Brazil; another question was obtained from a virtual learning environment and belongs to the realm of Philosophy: it proposes to argument on the differences of the four main eras of Philosophy.

The data sets of Biology and Geography were collected from an admission process at Federal University of Pará. For the corpus, the two questions with the highest frequency of answers were selected (130 Biology and 230 Geography answers): these answers have gone through a handheld digitization process, where only spelling errors were corrected and no adjustments to the grammatical aspects of the original text were made. The data set Philosophy was extracted from a virtual learning environment of the Federal University of Pará, from which 192 available answers were selected.

For each group of answers, a specific reference answer was used for comparison; for the Biology set, the reference response was a text produced by a specialist containing all possible and correct concepts; for the Geography and Philosophy sets, the reference answers were produced by concatenation of the best- scored answers in the human evaluation.

For the purpose of measuring the accuracy, the current study used the available scoring produced by human graders to the selected answers. For each response from both the Biology and the Geography sets two scores from two human graders were used. Likewise, for the Philosophy answers one score produced by one human grader was used.

A Table 1 presents, for each corpus, the quantity of answers; the minimum, maximum and average number of words per answer in each set; and the human vs. human accuracy.

Method

In this study, the automated evaluation based in text similarity techniques, performed comparisons of the students' answers against one reference response in each corpus. Prior to the comparison procedures, several phases of text preparation were necessary, for which we relied on the theoretical support of Burrows *et al*. (2015). These authors propose a five-stage pipeline architecture, where finishing up a previous phase means entering the next one, as demonstrated in Fig. 1.

While selecting the corpus (stage 1), spelling corrections were made, adjusting the vocabulary. In stage 2, filtration techniques were employed (conversion of uppercase letters into lowercase letters and deletion of graphic accentuation).

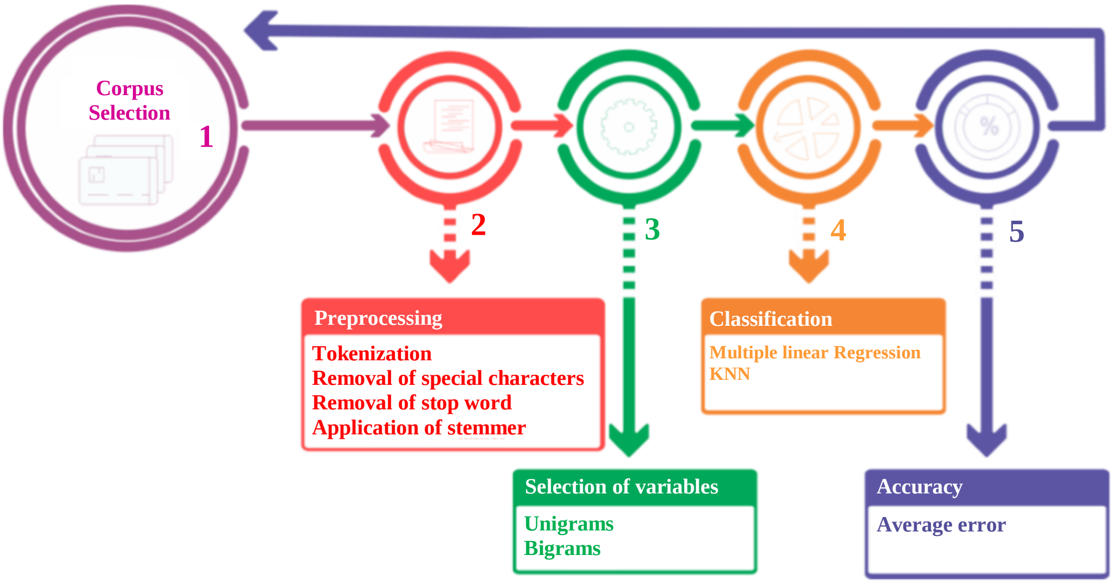


Fig. 1: Pipeline Architecture for automated evaluation of texts

After filtration, the answers were tokenized and three morphological processing techniques were used:

- Removal of Special Characters and Punctuation (+RCE)
- Removal of stop words (+RSW)
- Removal of suffixes (stemming) (+RSU)

The above techniques were combined in four different ways:

- Without preprocessing (-RCE, -RSW, -RSU)
- With removal of special characters (+RCE, -RSW, -RSU)
- With removal of special characters and stop words (+RCE, +RSW, -RSU)
- With removal of special characters and stop words, with the application of a stemmer (+RCE, +RSW, +RSU)

In stage 3, two vectors were generated: one for unigrams and the other for bigrams. It's worthwhile to remark that unigrams and bigrams possess distinct and complementary features. Unigrams are more frequently used to measure up similarity between texts, however, they do not take into account the sequence and organization of written texts, for this reason they are called "bag of words". Unlike unigrams, bigrams consider the writing order, even though they are less frequently used for measuring text similarity.

In stage 4, we explored similarity measures resulting from the combination of unigrams and bigrams from the sentence tokens (Table 2).

Table 2: Measures of similarity and frequency of terms

Set of terms	
Jaccard:	$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$
Overlap:	$\frac{\lvert A\cap B\rvert}{\min(\lvert A\rvert,\lvert B\rvert)}$
Dice:	$\frac{\lvert A\cap B\rvert}{\lvert A\rvert+\lvert B\rvert}$
Cosine:	$\frac{\lvert A\cap B\rvert}{\sqrt{\lvert A\rvert\cdot\lvert B\rvert}}$
Frequency of terms	
Euclidian Distance:	$\sqrt{\sum_i (a_i-b_i)^2}$
Cosine:	$\frac{< A,B >}{\lVert A\rVert\lVert B\rVert}$

Measures of similarity were based on the set of terms and frequency of words in the texts.

Still in stage 4, considering the vectors of similarity measurements and the scoring provided by the human graders as variables, it was possible to estimate a score for each response, replicating/simulating a human grader scoring by means of two classifications methods-multiple linear regression (Anderson, 2003) and KNN algorithm (K-nearest neighbors) (Cover and Hart, 2006).

In stage 5, to achieve the results, all the above- referred techniques have been combined, aiming at the best accuracy of the system vs. human. Complementarily, other questions, which had previously emerged in related works, were also examined:

Q1) Is it preferable to have one single reference response provided by a human specialist or to compose a reference response by using texts from the corpus itself? (Burrows *et al*., 2015)
Q2) What is the preprocessing impact over the final accuracy levels? (Salvatore *et al*., 2003; Burrows *et al*., 2015)
Q3) Regarding short answers, which are the most relevant measurements for term frequency? (Pribadi *et al*., 2017)
Q4) Does the combination of unigrams and bigrams produce good accuracy? (Magnini *et al*., 2005)
Q5) Which is the best classification method: linear regression or KNN?
Q6) Is the portability of the method across other linguistic domains possible? (Santos and Favero, 2015)

Results and Discussion

In this section, we present the main results achieved by means of the different combinations of morphological processing techniques, contrasted side by side to the human vs. human accuracy at Table 3; sequentially, Fig. 2(a), 2(b) and 2(c) show the accuracies obtained from a sample of 40 answers for corpus of Geography.

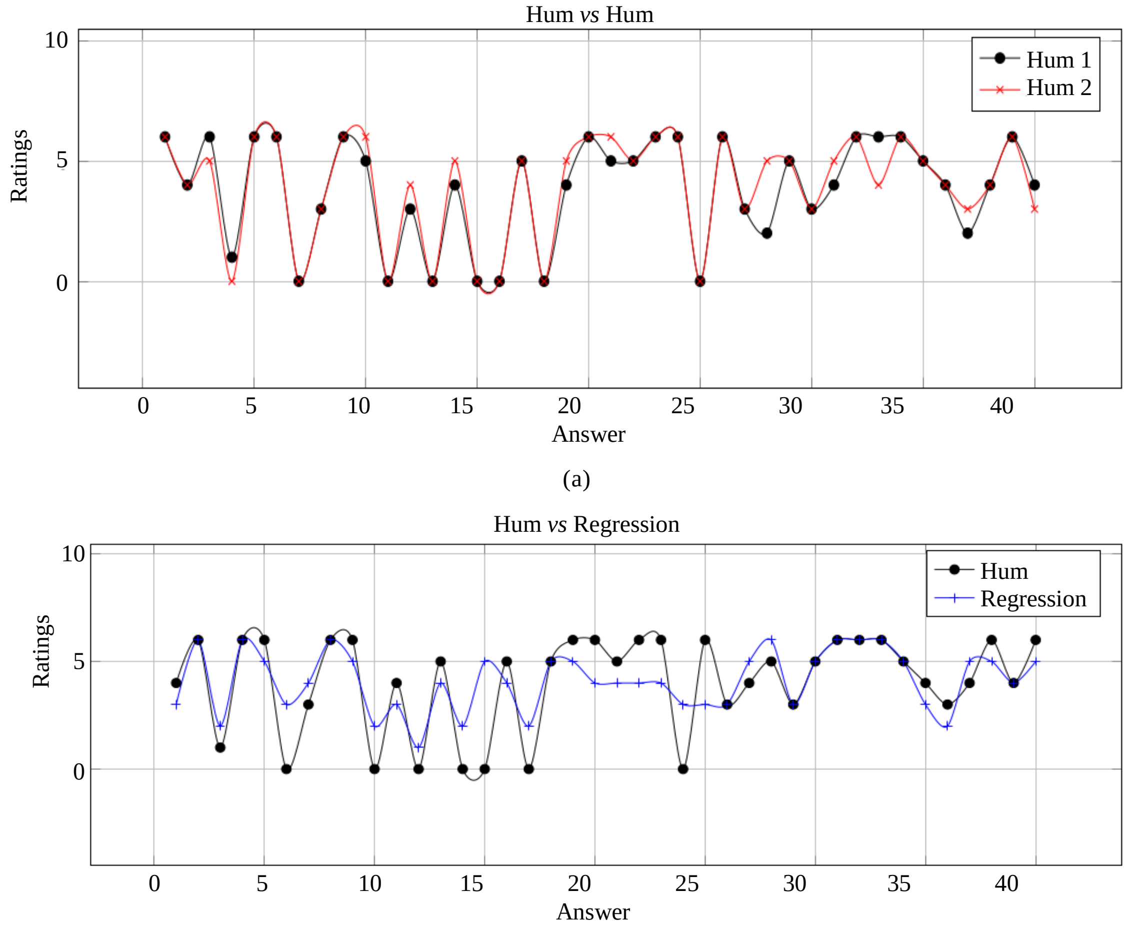
After presenting the results, we discuss and stress other findings stemming from the experiments, oriented by the six questions listed previously in Section Method.

Results from the Combination of Techniques Applied to Morphological Processing

The Table 3 presents the most substantial results of the Method contrasted side by side to the human vs. human results for each corpus. These outcomes derived from the experiments, which separately considered unigrams and bigrams and then jointly considered unigrams and bigrams by multiple linear regression.

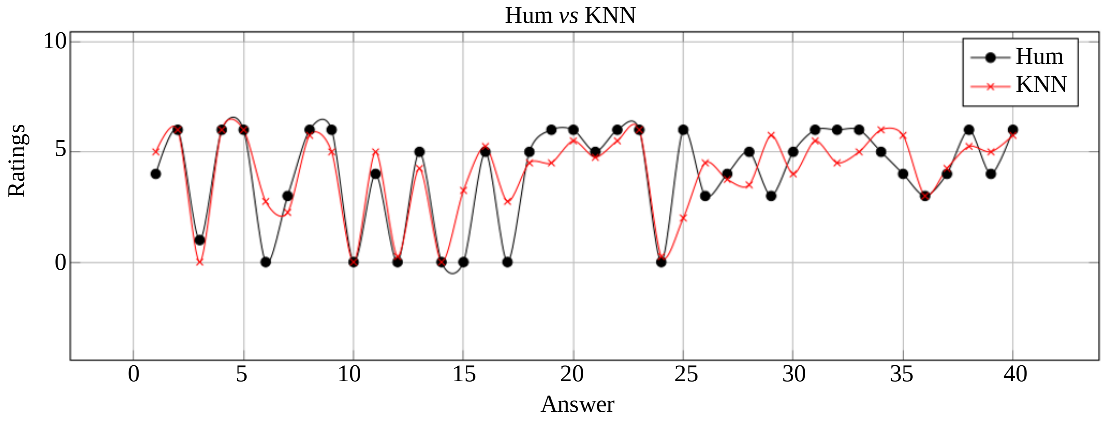
The above table shows that more preprocessing provided higher accuracy. For the Biology set 84.01, with KNN and for the Geography set 86.29, with regression, both with application of all preprocessing. However, for the Philosophy set, the highest index, 81.53, with KNN, was obtained with removal of special characters, punctuation and with removal of stop word.

Next we show the results of each specific examined corpus.



(a)

(b)



(c)

Fig. 2: An overview from a sample of 40 answers to the Geography question: (a) Human vs. Human (b) Human vs. Regression and (c) Human vs. KNN

Table 3: Comparison of human vs. human against Method results

	Biology		Geography		Philosophy	
	93.85		84.93		-	
Human vs. human	-----------------------------		---------------------------		---------------------------	
system vs. human	Regression	KNN	Regression	KNN	Regression	KNN
-RCE, -RSW, -RSU[a]	80.38	79.78	85.17	**85.07**	**81.59**	81.41
+RCE, -RSW, -RSU[b]	**80.77**	80.74	84.69	84.03	81.08	81.33
+RCE, +RSW, -RSU[c]	80.19	79.42	86.24	84.65	81.37	**81.53**
+RCE, +RSW, +RSU[d]	80.43	**84.01**	**86.29**	84.32	81.43	81.39

a. without preprocessing (-RCE, -RSW, -RSU)
b. with preprocessing: Removal of Special Characters and Punctuation (+RCE, -RSW, -RSU)
c. with preprocessing: Removal of stop words (+RCE, +RSW, -RSU)
d. with preprocessing: Removal of suffixes (stemming) (+RCE, +RSW, +RSU)

Results for the Biology Corpus

In this corpus, the best accuracy of system vs. human with linear regression was 80.77, while the best accuracy of system vs. human with KNN was 84.01. Between the two human graders the accuracy was 93.85.

Results for the Geography Corpus

The best accuracy of the Geography answers was 86.29, considering system vs. human with linear regression and 85.07 was the best accuracy considering system vs. human with KNN. Between the two human graders, the accuracy was 84.93.

Results for the Philosophy Corpus

The best accuracy of the Philosophy answers considering system vs. human with linear regression and system vs. human with KNN were, respectively, 81.59 and 81.53. As previously discussed, the Philosophy answers possessed one single human grader, the reason why it was not possible to measure up human vs. human accuracy.

Discussion

The experiments carried out indicate that a better accuracy is reached when using more preprocessing, but even though when unigrams and bigrams are combined. In view that bigrams consider word order in the texts, their use may help mitigate the “bag of words” phenomenon, whereas the use of unigrams alone may lead to increase vulnerability, as the use of repeated words may “deceive” the system. This evidence points to a reasonable, albeit temporary, answer to question Q4, presented earlier in section Method.

The best accuracy was verified in the Geography corpus, whose reference answer was the concatenation of the best-scored answers by human specialists. Considering question Q1 (see section Method), this finding suggests that it may be preferable to compose a reference answer by combining the best scored answers, also taking into account it shall offer a richer and better-adjusted vocabulary to the one more commonly used by students, as referred by Magnini *et al.* (2005).

However, it is important to observe that for the Philosophy corpus, whose reference response was also a product of concatenation of the highest scored answers, the accuracy levels were lower. This leads us to infer that the difference in system vs. human accuracy regarding each corpus (Biology, Geography, Philosophy) is likely to be related to the frequency of general vocabulary used, sometimes motivated by the type of the proposed question, or yet by the nature of the focused subject: with reference to our experiments, we highlight that even though the Biology question was one of conceptual nature, the reference response favored all possible response options; however, each student's answer, in compliance with the instructions, corresponded to only three options among the total universe of possibilities. For the Geography and Philosophy corpora, albeit both questions were of argumentative nature, it is possible to envisage a more predictable vocabulary in the Geography answers and a less predictable one in the Philosophy answers. This is probably accounted for by the nature of the topic approached in each question.

Regarding the preprocessing (Q2, section Method), the experiment indicates some positive evidence in the Biology and Geography tests. For example, the Biology test with KNN rises from 79.78 to 86.02; in the Geography with regression test rises from 85.17 to 86.29. Therefore, in these two questions, the best results were obtained with the use of preprocessing. However, for the answers of Philosophy, the preprocessing slightly decreased the accuracy: 81.59 without and 81.43 with.

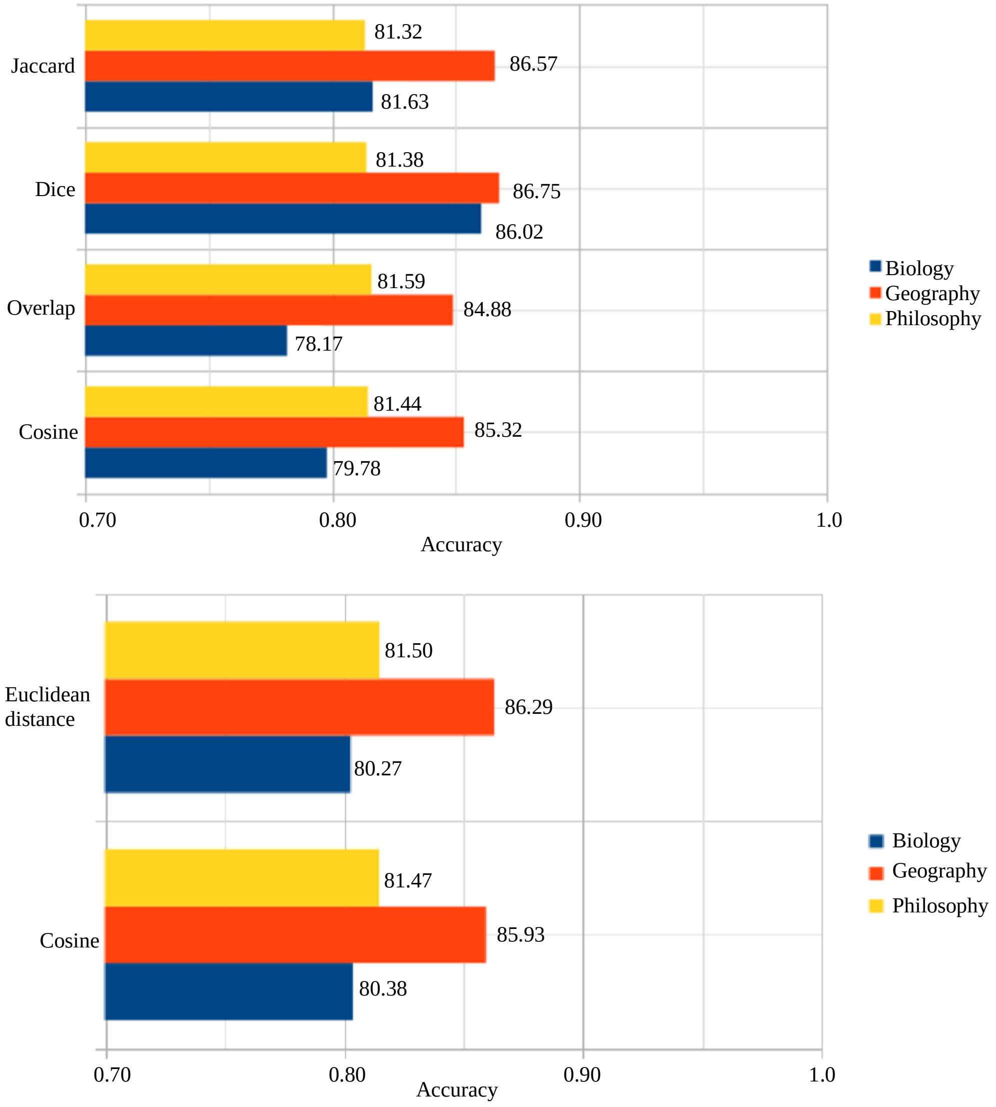


Fig. 3: Comparing the performances of the measures based on sets and the measures based on the terms' frequency

Table 4: Comparing the best performances between linear regression and KNN method

Corpus	Regression	KNN
Biology	80.77	**84.01**
Geography	**86.29**	85.07
Philosophy	**81.59**	81.53

Table 5: Comparing the best performances between linear regression and KNN method

	Accuracy		
SET	Minimum	Maximum	Average
20 tests human vs human	0.719	0.993	0.860
20 tests system vs human	0.607	0.989	0.840
67 tests human vs human	0.658	1.000	0.860
67 tests system vs human	0.662	1.000	0.850

The Fig. 3 contrast the four set based measures (Jaccard, overlap, dice and cosine) and the two terms frequency based measures (Euclidean distance and cosine vector). For the Biology, the Dice set based produced the best result, but for Geography the Euclidean distance, terms frequency based, produced the best result. In general, we have found very close results considering the different measurements. These results confirm the position advocated by Pribadi *et al*. (2017), who affirm that for short answers, set-based similarity measurements produce good results (Q3, section Method).

Table 4 shows that linear regression and KNN method present similar accuracy; however, it was necessary to create one regression model for each corpus, whereas the same KNN model could be used for all distinct corpora (Q5, section Method).

To test the potential for portability of the method across other linguistic domains, we performed experiments employing our approach to the corpus provided by (Burrows *et al*., 2015), with 2442 answers for 87 questions in the English language: the questions were categorized in two groups, the first of which with rating score ranging from 0 to 10 and the other one from 0 to 5. We used only the KNN method, with k = 4 and two measurements-Euclidian distance (unigrams and bigrams) and cosine (unigrams and bigrams). These questions have been scored by two human specialists with 0.86 accuracy between them. These results may be visualized bellow in Table 5.

The data show that on average, the system performed with very similar precision to that of human graders, pointing to the fact that the method portability is quite promising (Q6, section Method).

Conclusion

The purpose of this study was to develop a method of automated evaluation of short answers based on the text similarity measures. The goal was to find a combination of techniques with significant accuracy when compared to the accuracy verified between human specialists. Unigrams and bigrams were combined with multiple linear regression producing the best results. The KNN approach to classification proved promising for portability across linguistic domains. The results for the Biology corpus achieve the accuracy of 84.01 (system vs. human) against 93.85 (human vs. human). For the Geography corpus, the accuracy was 86.29 (system vs. human) against 84.93 (human vs. human) and for the Philosophy corpus, the accuracy was 81.59 (system vs. human). These results have demonstrated the potential of such technology for practical use in virtual learning environments.

Among the several experiments carried out along this study, we emphasize the one related to the reference answers, in which we could observe that the solution that turns the system more efficient may be either a answer given by a human specialist or the concatenation of the best answers among those that constitute the corpus.

As future work we intend to move in two directions: the inclusion of features of the technology of automatic evaluation of essays and, work with concepts similarity. The concepts are extracted from answers using deep machine learning and NLP, including triple extraction and Wikipedia, WordNet-based reasoning (Benomran and Ab Aziz, 2013; BenOmran and Aziz, 2018).

Acknowledgement

The authors thank the support of the Federal University of Pará (UFPA), PROPESP-UFPA. The authors thank the support of LabX team of Interdisciplinary Innovation – LabX-UFPA.

Author's Contributions

João A. dos Santos, Eloi Favero and Simone Negrão: Experiment, supervising, providing the research topic, reviewing and editing.

Silvério Sirotheau: Methodology, experiment and writing. This section should state the contributions made by each author in the preparation, development and publication of this manuscript.

Ethics

There are no ethical issues in publishing and giving open Access.

References

Anderson, T., 2003. An introduction to multivariate statistical analysis. Wiley Series in Probability and Statistics.

Attali, Y., B. Bridgeman and C. Trapani, 2010. Performance of a generic approach in automated essay scoring. J. Technol. Learning Assessment.

Benomran, A. and M. Ab Aziz, 2013. Automatic essay grading system for short answers in English language. J. Comput. Sci., 9: 1369-1382. DOI: 10.3844/jcssp.2013.1369.1382

BenOmran, A.M. and M.J.A. Aziz, 2018. New Information Content Glossary Relatedness (ICGR) approach for Short Text Similarity (STS) tasks. J. Comput. Sci., 15: 769-784. DOI: 10.3844/jcssp.2019.769.784

Burrows, S., I. Gurevych and B. Stein, 2015. The eras and trends of automatic short answer grading. Int. J. Artificial Intelligence Educ., 25: 60-117. DOI: 10.1007/s40593- 014-0026-8

Burstein, J., K. Kukich, S. Wolff, C. Lu and M. Chodorow *et al*., 1998. Automated scoring using a hybrid feature identification technique. Proceedings of the 36th Annual Meeting Association Computational Linguistics 17th International Conference Computational Linguistics, Aug. 10-14, Montreal, Quebec, Canada, pp: 206-210. DOI: 10.3115/980451.980879

Cover, T. and P. Hart, 2006. Nearest neighbor pattern classification. IEEE Trans. Inform. Theor., 13: 21-27. DOI: 10.1109/TIT.1967.1053964

Foltz, P.E., L.E. Streeter, K. Lochhaum and T.K. Landauer, 2013. Implementation and applications of the intelligent essay assessor. DOI: 10.4324/9780203122761.ch5

Gomaa, W.H. and A.A. Fahmy, 2012. Short answer grading using string similarity and corpus-based similarity. Int. J. Advanced Comput. Sci. Applic., DOI: 10.14569/IJACSA.2012.031119

Hearst, M.A., 2000. The debate on automated essay grading. IEEE Intelli. Syst. Their Applic., 15: 22-37. DOI: 10.1109/5254.889104

INEP., 2018. Instituto nacional de estudos e pesquisas educacionais anisio texeira.

Landauer, T., P. Foltz and D. Laham, 1998. An introduction to latent semantic analysis. Discourse Processes, 25: 259-284. DOI: 10.1080/01638539809545028

Magnini, B., P. Rodriguez, D. Perez, A. Gliozzo and E. Alfonseca *et al*., 2005. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. Revista Signos: Estudios De Linguıstica, 59: 325-343.

Mitchel, T., T. Russell, P. Broomhead and N. Aldridge, 2002. Towards robust computerised marking of free-text responses. Proceedings of the 6th CAA Conference, Loughborough: Loughborough University.

Mohler, M. and R. Mihalcea, 2009. Text-totext semantic similarity for automatic short answer grading. Proceedings of the 12th Conference European Chapter Association Computational Linguistics, Mar. 30, - Apr. 03, Athens, Greece, pp: 567-575.

Noorbehbahani, F. and A. Kardan, 2011. The automatic assessment of free text answers using a modified bleu algorithm. Comput. Educ., 56: 337-345. DOI: 10.1016/j.compedu.2010.07.013

Page, E.B., 1966. The imminence of... grading essays by computer. The Phi Delta Kappan, 47: 238-243. DOI: 10.2307/20371545

Pribadi, F.S., T.B. Adji and A.E. Permanasari, 2016. Automated short answer scoring using weighted cosine coefficient. Proceedings of the Conference e-Learning, e-Management e-Services, Oct. 10-12, IEEE Xplore press, Langkawi, Malaysia, pp: 70-74. DOI: 10.1109/IC3e.2016.8009042

Pribadi, F.S., T.B. Adji, A.E. Permanasari, A. Mulwinda and A.B. Utomo, 2017. Automatic short answer scoring using words overlapping methods. AIP Conference Proceedings, 1818: 020042. DOI: 10.1063/1.4976906

Rudner, L.M., V. Garcia and C. Welch, 2006. An evaluation of Intellimetric TM essay scoring system. J. Technol., Learning Assessment.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620. DOI: 10.1145/361219.361220

Salvatore, V., N. Francesca and A. Cucchiarelli, 2003. An overview of current research on automated essay grading. J. Inform. Technol. Educ., 2: 319-330. DOI: 10.28945/331

Santos, J.C.A. and E.L. Favero, 2015. Practical use of a Latent Semantic Analysis (LSA) model for automatic evaluation of written answers. J. Brazilian Comput. Society, 21: 21-22. DOI: 10.1186/s13173-015-0039-7

Shermis, M.D., C.M. Koch, E.B. Page, T.Z. Keith and S. Harrington, 2002. Trait ratings for automated essay grading. Educ. Psychol. Measurement, 62: 5-18. DOI: 10.1177/001316440206200101

Sukkarieh, J.G., S. Pulman and N. Raikes, 2003. Automarking: Using computational linguistics to score short. Free Text Responses.

Trusso Haley, D., P. Thomas, A. De Roeck and M. Petre, 2007. Seeing the whole picture: Evaluating automated assessment systems. Innovation Teaching Learning Inform. Comput. Sci., 6: 203-224. DOI: 10.11120/ital.2007.06040203

Zupanc, K. and Bosnić, Z, 2017. Automated essay evaluation with semantic analysis. Knowledge-Based Syst., 120: 118-132. DOI: 10.1016/j.knosys.2017.01.006