Santos and Favero *Journal of the Brazilian Computer Society* (2015) 21:21
DOI 10.1186/s13173-015-0039-7

Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers

João Carlos Alves dos Santos[1*] and Eloi Luiz Favero[1,2]

Abstract

This paper presents research of an application of a latent semantic analysis (LSA) model for the automatic evaluation of short answers (25 to 70 words) to open-ended questions. In order to reach a viable application of this LSA model, the research goals were as follows: (1) to develop robustness, (2) to increase accuracy, and (3) to widen portability. The methods consisted of the following tasks: firstly, the implementation of word bigrams; secondly, 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 taking into consideration the average of the words of the answers. The corpus was composed by 359 answers produced according to two questions from a Brazilian public university's entrance examination, which were previously scored by human evaluators. The results demonstrate that the experiments produced accuracy about 84.94 %, while the accuracy of the two human evaluators was about 84.93 %. In conclusion, it can be seen that the automatic evaluation technology shows that it is reaching a high level of efficiency.

Keywords: LSA, Automatic evaluation, Open-ended questions, Accuracy, Unigrams, Bigrams

Background

An automatic evaluation system is a computational technology used to analyze and rate written texts. Studies on this technology reveal how computers may be used to measure students' learning degree [1]. Though research on automatic evaluation of written answers has been going on since the 1960 [2–4], it has been only since the late 1990s that new models and methods of natural language processing have demonstrated higher accuracy level for practical applications [1, 5, 6], and more recent research shows accuracy closer to that of human evaluators [7].

The majority of research on automatic evaluation uses n-grams approaches within different applications [4, 8, 9]. More recently, regarding the automatic evaluation of written answers, there are some promising approaches that use the latent semantic analysis (LSA) model [4, 8, 10–12]. In these approaches, we find results ranging from 0.63 to 0.86, which are measurements of the correlation between scores attributed by LSA models and those attributed by human evaluators [10, 13, 14].

In this scenario, LSA models called bag-of-words (not concerned with word order) are adequate for many applications [15, 16]. More recent works investigate models which combine LSA with other techniques, such as word syntactic neighborhood analysis, a model that considers the arithmetic means between n-grams and LSA [4, 8, 17], or that combines *knn* algorithms with LSA [18]. In order to make the model more robust and possibly to enhance its accuracy, we work with bigrams of words because they preserve dependent relations between them that represent the sequence of words within the answers, thus surpassing the bag-of-words model.

Though some may argue that the use of bigrams with LSA is not viable for some applications due to size limitations on the initial matrices (the matrixes were around 1805 × 229), we find it possible to apply bigrams to this domain, since they are composed of short written answers (averaging 25 to 70 words) to open-ended questions from a higher education institution's exam. Furthermore, we only count those bigrams that appear at least

*Correspondence: jcas@ufpa.br
[1] Faculdade de Matemática, UFPa, Rua Augusto Corrêa s/n, Belém, Brazil
Full list of author information is available at the end of the article

in two answers, as to avoid a high number of null entries in the initial matrix.

In sum, our goal is to develop technology for an LSA system with more robustness, accuracy, and portability:

- *Robustness*: a traditional LSA model is not concerned with the order of the words within the text, thus being vulnerable to errors: an informed student may well deceive the system [10]. The use of bigram aims to make the system immune to this type of threat and therefore more viable.
- *Accuracy*: the correlation between an LSA system and a human evaluator may be high, it can be compared with correlation between two humans evaluator. This is a way to assert the efficiency of an LSA model, even though there still is no definitive research on the matter.
- *Portability*: the accuracy of an LSA model is largely dependent on the parameters of its applicability domain (type and size of corpus), making it difficult to apply successful experiments to new domains [10, 11, 19]. Calibration takes into account: preprocessing, local and global weighing, dimensionality, and the function of similarity.

The proposed LSA model estimates the scores of each answer with a six-step procedure: (1) preprocessing, (2) weighing, (3) singular value decomposition (SVD), (4) rating, (5) adjustments, and (6) accuracy. We then compare the LSA scores with those given by the human evaluators to calculate the accuracy between them.

An LSA model can be viewed as an improvement over a model based solely on *n*-grams because it has a more precise measure of similarity. LSA also captures contextual use of words, synonymy, etc. To demonstrate this fact, we tested a baseline model based on unigrams and bigrams combined with multiple linear regressions. This model reached an accuracy level of 78.74 %, while the LSA model reached 84.94 %. With these numbers, we can assert that this technology is efficient enough for practical application.

In this work, we propose an automatic evaluation approach for discursive questions. This system is acceptable to practical uses if its performance is similar to the human evaluators [14]. We believe that these may be innovative technologies for the virtual learning environment: we are employing this technique in a virtual learning environment of the Federal University of Pará, a college with 60,000 students from technical education through undergraduate to post-graduation. Further, this application is advantageous as an instant score feedback for large groups of students (hundreds or thousands), freeing the teacher from manual correction and enabling him or her to pay closer attention to lower scoring students.

This work is organized as an introduction plus six sections: the second presents a brief description of LSA, the third presents the corpus used in the work, the fourth presents the LSA model, the fifth presents the model's calibration process, the sixth presents a discussion of the results, and the seventh presents conclusions and future research.

A brief description of LSA

The first step is to create a term-document matrix A, of order m by n, m being the number of different words and n the number of texts in the corpus (this one represented by the column space of A). Each entry is weighed by a function that associates every word with its importance to the text from which it comes and within the whole corpus. The next step is to do the singular value decomposition (SVD) of A, revealing the architecture of correlations between the words in the texts. This way, we factor A into a product of three other matrices, in the form

$$A = USV^t,$$

where U and V are orthogonal matrices of orders m and n, respectively, and

$$S = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix},$$

D being a diagonal matrix with r singular values of A as its entries, and r is the rank[1] of A. After SVD, we approximate A with another matrix: we select the first k columns of U and V and the first k rows and columns of S to make a matrix $A_k = U_k S_k V_k^t$. The reason for this is that in the first columns of this matrix are the eigenvectors associated with the highest-magnitude singular values of A [20]. This approximating matrix A_k has the same dimensions of A, but it has a rank k generally much lower than r; A_k is the optimal approximation of A regarding quadratic norm [15]. This way, LSA transfers the analysis from our initial space to the lower-dimensional column space of A_k, called "semantic space", in which the rating step is made.

The corpus

Our research corpus is constituted of answers to two open-ended questions of the entrance examinations (vestibular exam) from the Federal University of Pará (Ufpa). In the period from 2004 to 2008, Ufpa decided to implement a form of selection that used tests that included open-ended questions from various disciplines, thus justifying that students would be better evaluated regarding skills and abilities related to college-level work. Each test consisted of 3 questions from 26 disciplines, and each student had to answer a single question from each discipline. During this period, more than 700,000 students took the tests resulting in more than 12 million written answers. From 1000 tests, we selected the two most often answered questions: the students only had to answer one third of the questions, so that the Biology question had been answered 130 times and the Geography question 229 times. During digitalization, orthography corrections were made to the original texts, but no changes in grammar were made.

The Biology question asked students to explain three concepts related to cell biology, whereas the Geography question asked the student to argue in defense of a given standpoint about the region's human and economic geography.

All 359 answers were previously scored by two human evaluators, receiving an integer score between 0 and 6. Each evaluator was unaware of the other's correction and score. There was also a check for discrepancy: if the two scores diverged by more than one point, a third evaluator would be given a score to be compared to the previous two scores.

The main point was to check the semantic similarity between each answer and a given "reference answer". That is why we consider the answer in our research. In the final stage of research, we used the answers given by a biologist as a gold answer. We chose to use the group of answers that had the highest scores attributed by human evaluators as the gold answer for the Geography question.

The LSA model

Our LSA model preprocesses the answers using unigrams and bigrams of words: it codes the rows of the initial matrix based on the occurrences of unigrams (or bigrams) and the columns show the answers; the reference answer is in the first column.

The model estimates the score of each answer with a six-step procedure:

1. Preprocessing: Making of the initial matrix: counts the unigrams and bigrams in each answer.
2. Weighing of the entries: a weight function expresses the importance of words in each answer and within the whole corpus.
3. SVD:
   (a) SVD calculation: the initial matrix is broken down into a product of three other matrices.
   (b) Reduction to semantic space: we empirically choose the dimension of semantic space.
4. Rating: each answer is compared to the reference answer.
5. Adjustments:
   (a) Penalty factor: based on the mean value and standard deviation of number of words per answer.
   (b) Re-rating: after applying the penalty factor, each answer is again compared to the reference.
6. Accuracy:
   (a) Error calculation: calculates the arithmetic mean of errors in each comparison.
   (b) Accuracy calculation:

$$\text{Accuracy} = \frac{6 - \text{Error mean}}{6} \times 100$$

Script executed several times, repeats steps 1 to 6 changing parameters and keeping the best configuration found.

Because the human evaluator score was an integer value between 0 and 6, it was necessary to categorize the scores of the LSA model: we partitioned the interval $[0, 1]$ into seven equal parts assigning each part the entire 0,1,2,3,4,5, and 6, respectively. We then compare the LSA scores with those given by the human evaluators to calculate the accuracy between them: the value absolute of the difference between the scores is the percent error. Adding all the percent errors and dividing by the number of answers, we get *error mean*.

The model is executed a number of times to find the optimal calibration for the parameters that influence the model's efficiency. More than 60,000 executions were performed. Step 1 was programmed in JAVA, while the other steps were programmed in MATLAB. The next session shows how the calibration of the LSA model was done.

Methods to calibrating the model

Calibration was performed during the running of the experiment: it was feasible to use an approach that found the best possible values for each parameter, for the best accuracy. On preprocessing, variations were considered for the entries of A:

a) Counting all the stop words
b) Removing all the stop words
c) Removing all the stop word plus a stemming process

This allowed constructing of six distinct initial matrices for each set of answers.

Research agrees that unigrams with removal of stop words and stemming provide the best results for LSA [14]. Nevertheless, when we use bigrams, the stop words take on the role of "function words"; as so, we opted to count the stop words.

On preprocessing, we only kept those bigrams that appeared in at least two answers, as to avoid a large number of zero entries.

During the weighing step, the matrix A undergoes a preliminary transformation called weight function, defined as the product of a number that represents a local weighing by another that represents a global weighing. The weighing scheme used was an application of the TermFrequency(tf) vs. InverseDocumentFrequency(idf) transform, the same scheme used by Dumais and other [1, 19, 21–25].

On the SVD step, we chose the optimal size of semantic space; this step has had the most notable impact on our final results. Some suggestions can be found in literature regarding the choice of the dimensionnality k [22, 23]; however, there is no consensus. The value of k is a function of the size of the matrix they are using. We favor the brute force method: to vary k from 1 to the total number of answers, thus picking the optimal value. The best results were obtained for values of k between 2 and 8, probably due to the fact that the SVD orders the eigenvectors of largest to smallest magnitude of eigenvalue associated. Other works consider the dimensionality as 94, 100, 200, or 300 [10, 13, 26, 27].

On the classification step, we estimate the similarity between the reference answer and the rest of the answers. The vector space approach was used: each answer is converted into a vector and, through calculations with these vectors, we were able to provide a value for the similarity between answers. The best results were obtained by using the cosine of the angle between two vectors, although the Pearson correlation has provided similar results. Some works used an approach combining Euclidian distance and cosine [23, 28]. We ran again some experiments considering only the Euclidian distance, but there was no improvement in the results.

In our experiments, we encountered as a problem the fact of the automatic evaluator that assigned a high score to a response that contained only a few words. It happens when we have a big vector reference response and a small response vector with a small angle between both, what results in a high value to the cosine measure. The works of Olmos et al. [28] and Jorge-Botana et al. [23] corrected this problem combining the cosine measure with the Euclidian distance. We opted to implement an alternative technique which penalizes short responses. In future works, we intend to investigate which approach is better to correct the problem with short responses.

Results and discussion

This study had four main objectives: to create a model of co-occurrence of unigrams and bigrams, to combine bigrams and LSA, to adjust the LSA scores based on the number of words per answer, and to compare LSA-attributed score distribution with that of human evaluators.

Model of co-occurrence of unigrams and bigrams

Considering the successful *n*-grams based on research found in literature [1, 4, 29, 30], we decided to utilize an *n*-grams model as our baseline with three scenarios: only unigrams, only bigrams, and unigrams combined with bigrams.

The baseline model measures the similarity of answers with the reference answer, taking into account the number of unigrams or bigrams which are found in common in both answers. For all cases, we used linear regression to approximate the scores with those of human specialists as done in Nikos et al. [29] using the metric mutual information between the words of answers (corpus-based metric) in a *n*-gram regression model that obtained a correlation of 0.62 when comparing two sentences.

Those values from baseline model were used as reference for the LSA model's performance.

Biology question

We note graphically in Fig. 1 similarities between the *n*-gram models' and human evaluators' scores ranging from 3 to 5. There's a gap of more than 2 points for scores below the 2 points. We also note small differences around the 6 points. The graph shows that the behavior of unigrams and bigrams was the same.

The indexes of accuracy considering only *unigrams* and *bigrams* were 78.5 and 75.37 %, respectively. To combine unigrams and bigrams in a single variable, we used a multiple linear regression model and for this variable, the index was 78.93 %. For both human evaluators, this index was 93.94 %.

Geography question

In Fig. 2, we observe that the *n*-gram models diverge by 1 point from the human evaluators in the lower scores, are alike in the mid ones, and differ a little up to scores of 5 points. The graph also shows that the unigrams and bigrams worked practically the same.

The indexes of accuracy considering only unigrams and bigrams were 83.89 and 83.77 %, respectively. Combining unigrams and bigrams through a multiple linear regression model, the index was 83.95 % and that of the human evaluators was 84.93 %.

The model baseline reached an accuracy index close to that of the human evaluators for the Geography question, but the same did not happen with the Biology question.

LSA model

The similarity estimation on the LSA model is made after applying the penalty factor; it considers every combination of preprocessing.

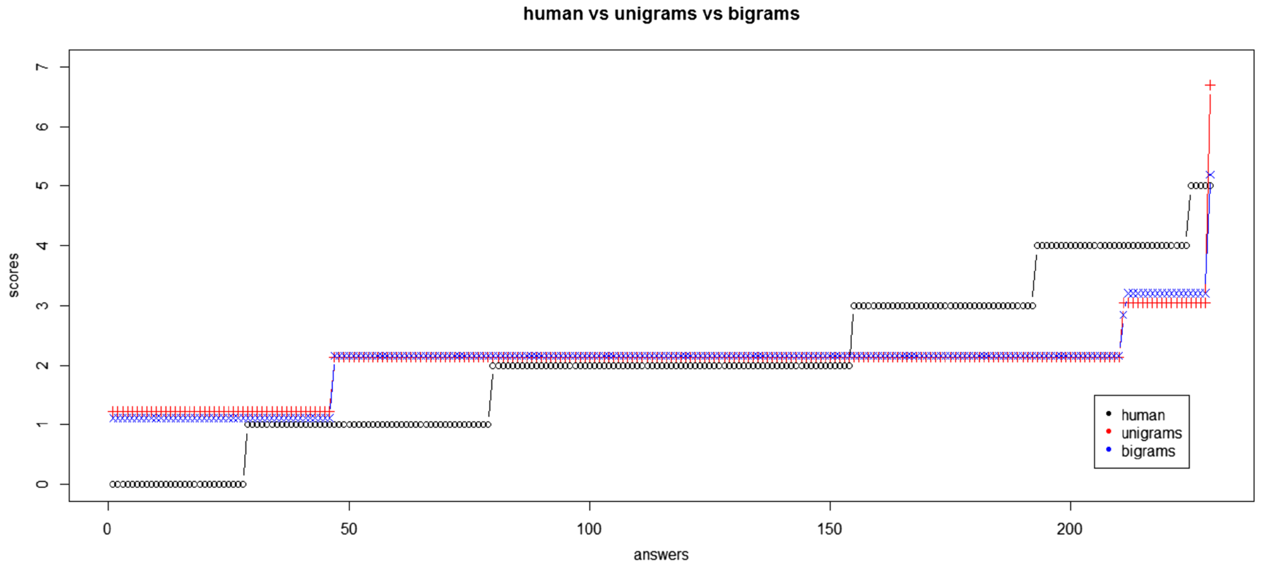

Fig. 1 Model performance of model baseline for the Biology question. Scores attributed to the Biology question's answers by human evaluators and model baseline (unigrams and bigrams)

Biology question

We notice in Fig. 3 a small difference between the human evaluators' curve and the LSA model's for scores 1 point below, but we can see an important correlation between the two on the rest of the graph.

The indexes of accuracy for unigrams with LSA and bigrams with LSA were 83.07 and 83.46 %, respectively; the index of human evaluators was 84.93 %. No multiple linear regression model was applied.

To better interpret the results, we compared the score distribution and perform a comparison test of averages of the LSA model and human evaluators for the Biology question's answers.

We notice in Table 1 that the LSA model reached 71.54 % between coincident scores and with a difference of 1 point.

The distribution of scores reveals that the LSA model assigned five scores of 0 and the human evaluator 16, justifying the graphical difference for this score. In the scores

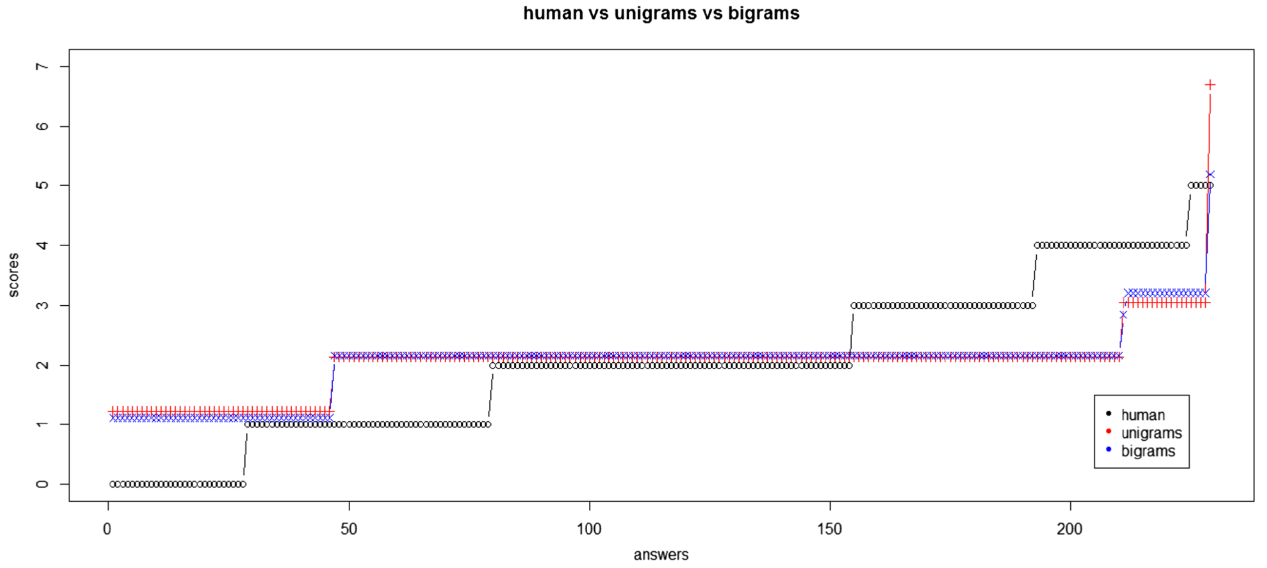

Fig. 2 Model performance of model baseline for the Geography question. Scores attributed to the Geography question's answers by human evaluators and model baseline (unigrams and bigrams)

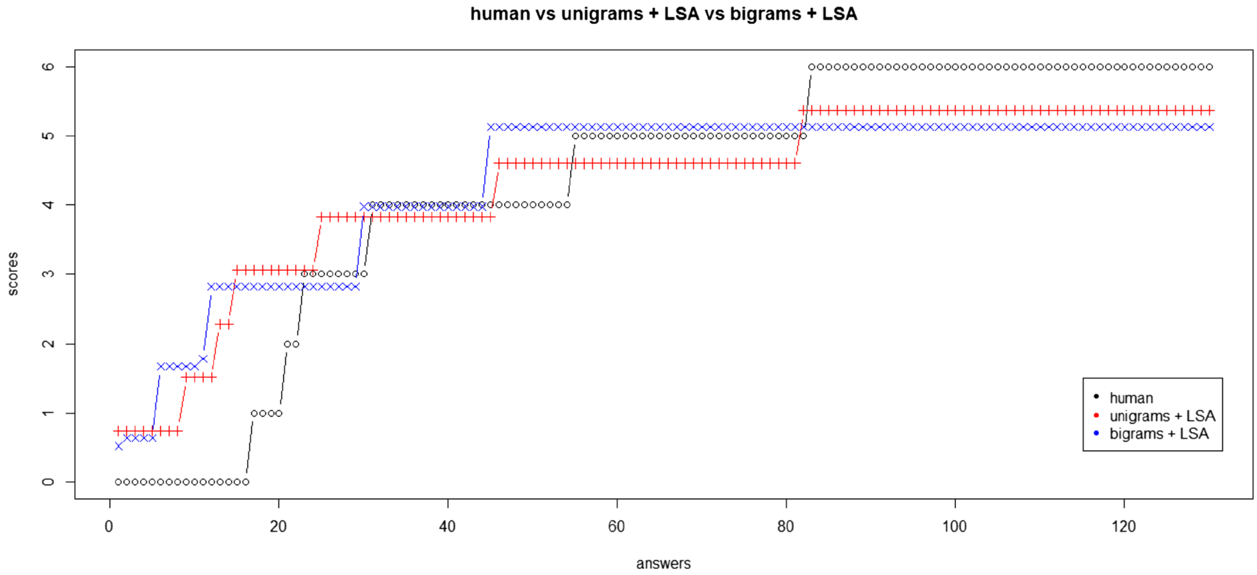

Fig. 3 Model performance of model LSA the Biology question. Scores attributed to the Biology question's answers by human evaluators and model LSA (unigrams and bigrams)

1 to 6, the LSA model assigned 125 scores and the human evaluator 114 scores, justifying the proximity of the graphics in this interval of scores. The LSA model overestimated the high score 6. The difference between scores was not statistically significant.

Geography question

We observe in Fig. 4 a difference of 1 point between the human evaluators and LSA models curves below the 1 point, considering all score scales, great similarity until scores of 4 points, and then another gap. The graph shows similarity between the LSA models on all scores.

The indexes of accuracy for unigrams with LSA and bigrams with LSA were 84.94 and 84.15 %, respectively, and the index of the human evaluators was 84.93 %. No multiple linear regression model was applied.

To better interpret the results, we compared the score distribution and perform a comparison test of averages of

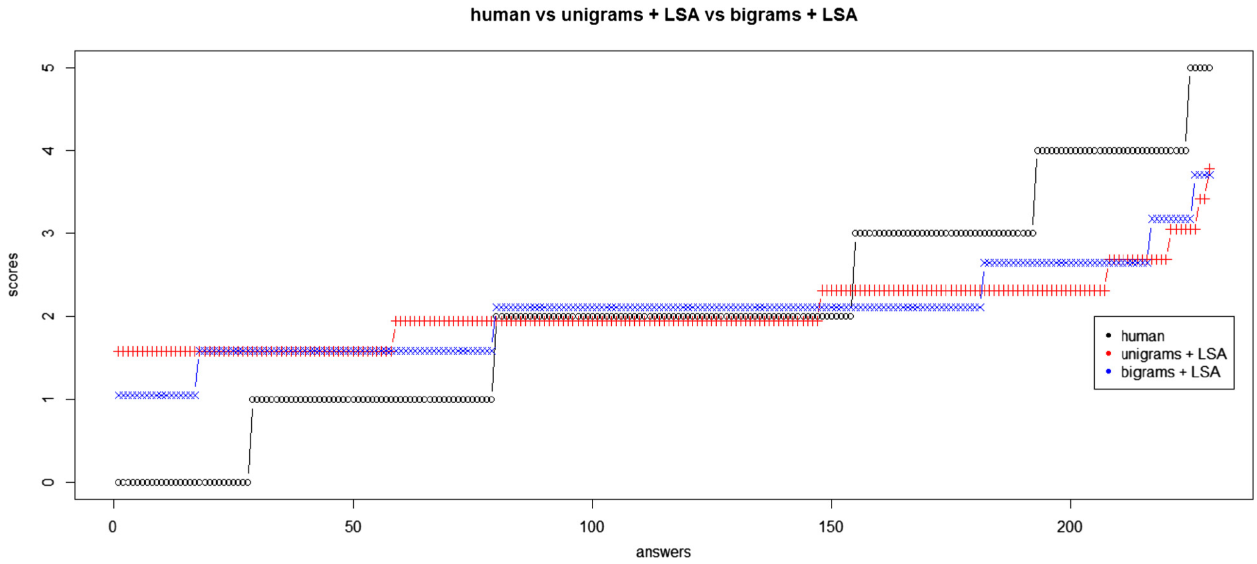

Fig. 4 Model performance of model LSA the Geography question. Scores attributed to the Geography question's answers by human evaluators, unigrams with LSA, and bigrams with LSA

Table 1 Absolute frequency LSA × human

		Human							Total	Percent (%)
		0	1	2	3	4	5	6		
LSA	0	3	1	0	0	0	1	0	5	
	1	2	1	0	0	0	0	0	3	
	2	3	1	0	2	1	0	0	7	
	3	3	0	1	4	2	2	3	15	
	4	0	0	0	2	7	4	6	19	
	5	2	0	0	0	2	5	4	13	
	6	3	1	1	0	12	16	35	68	
Total		16	4	2	8	24	28	48	130	
Number of coincidences									55	42.31
Number of close values									38	29.23

the LSA model and human evaluators for the Geography question's answers.

We notice in Table 2 that the LSA model reached 79.91 % between coincident scores and with a difference of 1 point.

The distribution of scores reveals that the LSA model assigned half of scores to the human evaluator, justifying the graphical difference for this score. The LSA model underestimated the minimum score 0. There is a relative equivalence in the distribution of score among the other intervals. The difference between scores was not statistically significant.

Baseline model vs. LSA model

Biology question

The indices of accuracy obtained by LSA model were better: 78.5 % vs. 83.07 % in the unigram approach and 75.37 % vs. 83.46 % in the bigram approach. The numbers show that the LSA model is a refinement of the baseline model. A possible explanation resides in the fact that the *n*-gram models measure an answer's similarity to a reference answer considering their shared unigrams and bigrams, but not the force of connection between the words. The difference between the LSA model and the human evaluator stayed above 10 %, which still is not desirable. A possible reason for this is the fact that the answers of Biology were compared to a single answer given by a human expert.

Table 2 Absolute frequency LSA × Human

		Human							Total	Percent (%)
		0	1	2	3	4	5	6		
LSA	0	10	2	2	0	0	0	0	14	
	1	11	24	12	6	3	1	0	57	
	2	4	13	33	6	9	0	0	65	
	3	3	7	16	14	5	1	0	46	
	4	0	4	12	11	8	1	0	36	
	5	0	1	0	1	5	1	0	8	
	6	0	0	0	0	2	1	0	3	
Total		28	51	75	38	32	5	0	229	
Number of coincidences									100	43.67
Number of close values									83	36.24

Geography question

The numbers show that the LSA model and the baseline model performed equally: 84.94 % vs. 83.89 % in the unigram approach and 84.15 % vs. 83.77 % in the bigram approach. This is mostly likely due to the fact that the reference answer is a concatenation of the three highest scoring answers—as attributed by the human evaluators. The distribution reveals the model attributed scores ranging from 1 to 4 for 37 answers, while the human evaluators count 38 answers in the same interval. The model placed 85 answers in the 5- to 6-point interval, while the human evaluators have only done so for 76. This indicates that the model has overestimated the highest scores, but nevertheless, is similar in the lower values.

It has been observed that the baseline model did not have the same performance in the two sets of answer, while the LSA model had the same performance and provided evidence supporting its robustness.

Conclusions

Our focus was to develop a technology allowing practical use of an LSA model for automatic evaluation of written answers to open-ended questions. We worked in four directions: (1) to create a co-occurrence model of unigrams and bigrams and use it as the baseline for the LSA model—it was verified that LSA model was superior in both approaches; (2) to combine bigrams with LSA to test its performance, since traditional models only use unigrams. We observed that this combination produced better results with the conceptual answers, but did not have any significant impact on the evaluation of the argumentative answers. In both sets of answers, the use of bigrams did not bring any improvement compared to unigrams; (3) to adjust the LSA scores based on the number of words per answer. The implementation of this penalty factor caused an enhancement of 8 to 10 % on the accuracy indexes; (4) to analyze the distribution of the LSA attributed scores against the human evaluators' scores, having found 79 % of coincident or close answers. The results show that LSA models can be used to refine results from methods based solely on *n*-grams, with the best outcomes depending on parameters calibration and on the applicability domain. The experiments provided an accuracy index of 84.94 % compared to human evaluators, whereas the accuracy between those evaluators was of 84.93 %. In this study's domain, the technology presented results close to human evaluators' results, showing actual application potential in automatic evaluation systems in virtual learning environments.

Endnote

[1] The rank of a matrix M is the dimension of row or column space of M

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Federal University of Pará's Selechive Process Centre (SPCE) for providing the data.

Author details

[1] Faculdade de Matemática, UFPa, Rua Augusto Corrêa s/n, Belém, Brazil.
[2] Faculdade de Computação, UFPa, Rua Augusto Corrêa s/n, Belém, Brazil.

Received: 6 May 2014 Accepted: 16 October 2015
Published online: 21 November 2015

References

1. Marinrez DP, Alfonseca E, Rodriguez P, Gliozzo A, Strapparava C, Magnini B (2005) About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. Revista Signos 38:325–43
2. Baker FB (1962) Information retriviel based upon latent class analysis. J AMC 9:521–1
3. Hearst MA (2000) The debate on automated essay grading. IEEE Intell Syst 15:22–37
4. Noorbehbahani F, Kardan AA (2011) The automatic assessment of free text answers using a modified bleu algorithm. Comput Educ 56:337–45
5. Lifchitz A, Larose SJ, Denhiere G (2009) Effect of tuned parameters on an LSA multiple choice questions answering model. Behav Res Methods 41:1201–1209
6. Graesser AC, Rus V, D'Mello SK, Jackson GT (2008) AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner, vol. 1. In: Robinson DH, Schraw G (eds). Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning. Information Age Publishing. pp 95–125
7. Burstein J, Chodorow M, Leacock C (2003) Criterion SM: online essay evaluation: an application for automated evaluation of student essays. In: Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence. Association for the Advancemet of Artificial Intelligence
8. He Y, Hui SC, Quan TT (2009) Automatic summary assessment for intelligent tutoring systems. Comput Educ 53:890–9
9. Chu-Carroll J, Carpenteru B (1999) Vector-based natural language call routing. J Comput Linguist 25:361–88
10. Landauer TK, Foltz PW, Laham D (1999) An introduction to latent semantic analysis. Discourse Process 25:259–84
11. Bestgen Y (2006) Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-hastings, and Moore. Comput Ling 32:5–12
12. Magliano JP, Graesser AC (2012) Computer-based assessment of student-constructed responses. Behav Res Methods 44:608–21
13. Wolfe MBW, Schreiner ME, Rehder B, Laham D (1998) Learning from text: matching readers and texts by latent semantic analysis. Discourse Process 25:309–36
14. Haley DT, Thomas P, Roeck AD, Petre M (2007) Seeing the whole picture: evaluating automated assessment systems. ITALICS 6:1473–1507
15. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inform Sci 41:391–407
16. McNamara DS (2011) Computatioinal methods to extract meaning from text and advance theories of human cognition. Topics Cognitive Sci 3:3–17
17. Kanejiya D, Kumar A, Prasad S (2003) Automatic evaluation of students answers using syntactically enhanced LSA. In: Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2. Association for Computational Linguistics
18. Layfield C (2012) With LSA size does matter. In: Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation, EMS 2012, Malta. pp 127–31
19. Dumais ST (1991) Improving the retrieval of information from external sources. Behav Res Methods Instrum Comput 23:229–36
20. Lay DC (2011) Linear algebra and its applications. Fourth edition, Vol. 1. Pearson, Addison-Wesley
21. Nakov P (2000) Chapter 15: Getting better results with latent semantic indexing. In: Proceedings of the Students Prenetations at ESSLLI-2000
22. Wild F, Stahl C, Stermsek G, Penya Y, Neumann G (2005) Factors influencing effectiveness in automated essay scoring with LSA. In: Proceeding of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology
23. Jorge-Botana G, Leon JA, Olmos R, Escudero I (2010) Latent semantic analysis parameters for essay evaluation using small-scale corpora. J Quant Linguist 17:1–29
24. Zen K, Iskandar DNFA, Linang O (2011) Using latent semantic analysis for automated grading programming assignments. In: International Conference on Semantic Technology and Information Retrieval. IEEE
25. Reafat MM, Ewees AA, Eisa MM, Ab Sallam A (2012) Automated assessment of students arabic free-text answers. Int J Cooperative Inform Syst 12:213–222
26. Foltz PW, Laham D, Landauer TK (1999) The intelligent essay assessor: applications to educacional technology. Interactive Multimedia Education Journal of Computer enhanced learning 1(2). http://www.bibsonomy.org/bib/bibtex/264bd31f5a77a7ab667664219d9f66acb/quesada
27. Wiemer-Hastings P, Wiemer-Hastings K, Graesser A (1999) Improving an intelligent tutor's comprehension of students with latent semantic analysis. In: Artificial Intelligence in Education. IOS Press
28. Olmos R, Leon JA, Jorge-Botana G, Escudero I (2009) New algorithms assessing short summaries in expository texts using latent semantic analysis. Behav Res Methods 41:944–50
29. Malandrakis N, Iosif E, Potamianos A (2012) DeepPurple: Estimating Sentence Semantic Similarity using N-gram Regression Models and Web Snippets. Association for Computational Linguistics 565–570. http://dl.acm.org/citation.cfm?id=2387731
30. Islam MM, Hoque ASML (2012) Automated essay scoring using generalized latent semantic. J Comput 7:616–26