Portuguese Automatic Short Answer Grading

Lucas B. Galhardi[1], Cinthyan R. S. C. Barbosa[1], Rodrigo C. Thom de Souza[2], Jacques D. Brancher[1]

[1]Graduate Program in Computer Science – Londrina State University (UEL)
10.011 - 86057-970 - Londrina, PR - Brazil

[2]Natural and Scientific Computing Research Group – Federal University of Paraná (UFPR)
86900-000 - Jandaia do Sul, PR - Brazil

{lucasbgalhardi,cinthyan,jacques}@uel.br, thom@ufpr.br

Abstract. *Automatic Short Answer Grading is the study field that addresses the assessment of students' answers to questions in natural language. Besides length, it differs from automatic essay grading by focusing on the evaluation of content instead of answer's style. The grading of the answers is generally seen as a typical classification supervised learning. Many works have been recently developed, but most of them deal with data in the English language. In this paper, we present a new Portuguese dataset and system for automatic short answer grading. The data was collected with the participation of 13 teachers, 12 undergraduate students and 245 elementary school students. Results achieved 69% accuracy in four-class classification and 85% on binary classification.*

1. Introduction

Evaluations are employed to demonstrate acquired knowledge in the students' learning process. Despite the importance of evaluation, teachers usually find the task of assessing the respondents' answers time-consuming. Also, students may have to wait for a long time to receive feedback on their responses and, when they finally get it, the grade can be different from another classmate's, who has given a very similar answer [Santos et al. 2012, Passero et al. 2016].

The computer-based assessment addresses these issues and improves other aspects of learning by automating the evaluation process. Evaluations are often composed of recall and recognition type of questions, which are in different levels of the learning depth. Recognition kind seeks to test the respondent's ability to organize or identify some specific information, while for recall, respondents need to remember external knowledge and write their own answers. Automatic grading is a solved problem for recognition questions, but it is an open problem and research subject for the recall kind [Burrows et al. 2015].

The automatic assessment bring some benefits such as formalization of correction criteria, delivery of faster feedback to both teacher and student and the better use of teachers' time. [Liu et al. 2016].

There are many different types of questions that can be required from students. Short answers are the focus of interest in this work. They can range from one sentence to one paragraph (few sentences), must be written in some natural language and recalls to external knowledge outside the question statement. Moreover, the evaluation is made with a focus on the content rather than style. This research field is defined in

DOI: 10.5753/cbie.sbie.2018.1373

[Burrows et al. 2015] as Automatic Short Answer Grading (ASAG) and used in later works [Roy et al. 2016, Zhang et al. 2016]. It consists in automatically assessing short natural language responses using computational methods.

The goal of this work is to present a new Portuguese dataset and system for automatic short answer grading. The data was collected through a web application specifically designed to work in the short answers context. We then implemented our method to automatically grade the collected students' answers based on techniques from specialized literature and reported the results.

The remaining of this paper is organized as follows. Section 2 reviews related works and give an overview of Portuguese researches. Section 3 exposes the data used in this work, concerning its collection, conditions, treatment and characteristics. In Section 4 the proposed method for this work is described. Then, in Section 5 the experiments in the data are reported and discussed. Finally, Section 6 presents our final considerations.

2. Related Work

Computer-based grading of students knowledge started long ago, with the work of [Page 1966]. Since then, hardware and software technology evolved so much that today distance education through virtual learning environments has become common. Considering this scenario, the need for automatic assessments is constantly increasing.

In [Burrows et al. 2015] the Automatic Short Answer Grading research field is formally defined, stating the differences from automatic essay grading and reviewing 35 different ASAG systems from 1996 to 2013, showing that research on the field started in a not unified front, with each research reporting experiments in its own private data.

The scenario in ASAG research starts to change in 2011 when the first three public datasets were released (the Texas[1] and the CREE and CREG datasets from the CoMic project[2]). These datasets opened possibilities for future works fair comparisons and began the "Evaluation Era" as stated by [Burrows et al. 2015].

Following towards the evaluation trend, in 2012 and 2013 two competitions challenged participants to develop ASAG systems for three new datasets. In 2012, Kaggle, a data science competition website, released the Automated Student Assessment Prize: Short Answer Scoring[3]. It provided competitors and future researchers with about 22000 human graded answers to 10 interdisciplinary questions.

Thereafter, the 2013 Semantic Evaluation organizers released in their 7th task the Joint Student Response Analysis and Eighth Recognizing Textual Entailment Challenge [Dzikovska et al. 2013]. It presented researchers with two datasets about electronics and general science with more than 100 questions and thousands of graded answers.

The availability of these six public datasets from 2011 to 2013 stimulated research on the ASAG field beyond their original authors and competitors as between 2014 and 2017 new works were developed using these datasets (e.g. [Higgins 2014, Ramachandran et al. 2015, Roy et al. 2016, Riordan et al. 2017], just to name a few).

[1] web.eecs.umich.edu/ mihalcea/downloads.html

[2] www.uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-833/section-a-context/a4-meurers/software-resources-and-corpora.html

[3] www.kaggle.com/c/asap-sas

2.1. Portuguese Related Work

When searching for Portuguese ASAG systems, seven researches were found. They can be divided into three groups. The first is by works that employed manually crafted linguistic, syntactic or fuzzy rules to match student answers to the expected content. This kind of automatic grading does not need a large amount of data, but the grading system is restricted to the data. They worked with 1, 5 and 30 questions and 68, 15 and 300 student's answers in total, respectively [Salton et al. 2013, Vilela et al. 2012, Flores et al. 2014].

The second group of works consists of those who used ngrams and similarity metrics (with expected answers) to grade students answers. These researches accomplished similarity by using lexical or semantic approaches, in the form of string-matching, Latent Semantic Analysis and WordNet similarity [Santos et al. 2012, Ávila and Soares 2013, Passero et al. 2016]. These similarity techniques are grouped in a survey [Vijaymeena and Kavitha 2016] and later explained in this work (see Section 4). This group of works also uses data between 1 and 13 questions and 76 and 359 student's answers in total.

The last group is the most similar and related to our own research. It consists of only one work [Figueira et al. 2013] that employed supervised machine learning techniques to grade the answers. Authors collected data from an existing virtual learning environment, in which they extracted 17000 answers to 31 questions. As features, they used ngrams and lexical similarity values obtained from some metrics.

3. Portuguese ASAG Data

The data addressed in this work (publicly available at[4]) is originated from a web system[5] specifically designed to handle short answers in the learning context, directed to be used by teachers and students. The system manages exams, questions, reference answers from teachers, students answers and their correspondent grades. A usual workflow of the system consists in the creation of an exam and its composing questions. Then, students can select the exam and answer its questions. Finally, grades are assigned for each answer by a teacher and by the automatic grading system.

Until this moment, the system has one exam with 15 questions created together by five elementary school biology teachers. The subject matter addressed by the questions consists mainly of human body topics. Some examples are: *"Explique o mecanismo de inspiração e de expiração do ar no corpo humano:"* and *"Quais são as diferenças entre veias e artérias?"*. For each question, between two and four reference answers were also created by the teachers, alongside with between three and six keywords.

The recorded exam was then applied to 245 elementary school students (8th and 9th grades, about 12-14 years old). The application was made with the supervision of their teachers and each student had to come up with its own answers to the questions. Most students answered directly in the web application but, in some schools with fewer conditions, the application was made in paper and then transcribed (rigorously, including spelling errors, spaces, accentuation, etc) to the system. No answers were left in blank as students were instructed to do so, even if they did not know the answer. This was done

[4]www.kaggle.com/lucasbgalhardi/pt-asag-2018
[5]www.autoavaliadorcir.com/about/

in order to collect the maximum number of answers, being correct or not. The number of sentences in answers ranges from 1 to 8 (1.5 in average) and the number of words varies from 1 to 136 (12.7 in average).

In possession of the 3675 answers (15x245), 12 undergraduate biology students from the final years of college evaluated the answers considering a predefined scale. Graders assigned one of four possible grades to each answer: **0 -** when the answer is wrong, **1 -** if the answer has something correct but it is still mostly wrong or incomplete, **2 -** if the answer is correct but has some wrong detail or missing important content and **3 -** if the answer is mostly correct, with the important points presented.

Completely equal answers were removed as they consisted in duplicated data, impairing machine learning algorithms. The difference between before and after this process can be seen in the label distribution histogram presented in Figure 1. The complete label distribution per question is presented in Table 1 (already considering removed duplicates). Questions 7, 8, 9 and 10 (gray marked in the table) are highly unbalanced and, therefore, were removed from our experiments as they do not contribute to performance analysis. This is due to machine learning algorithms that generally do not work as expected on these conditions of highly unbalanced data.

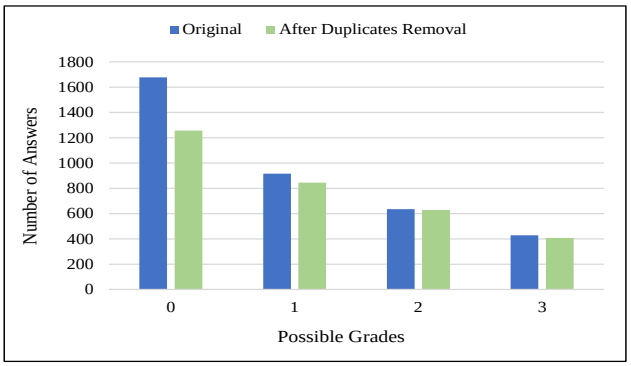

Figure 1. Labels Distribution.

Table 1. Labels Distribution.

Q_ID	0	1	2	3	Q_ID	0	1	2	3
1	59	44	58	72	**9**	149	18	10	11
2	60	77	81	20	**10**	144	11	16	5
3	85	105	21	8	**11**	106	33	78	12
4	77	49	26	22	**12**	86	36	21	61
5	82	40	40	46	**13**	54	72	48	40
6	65	44	61	40	**14**	37	65	101	29
7	110	92	6	1	**15**	85	50	38	16
8	72	120	18	3	**Sum**	1271	856	623	386

Only 22% of the answers received more than one grade. From this subset, the inter-rater reliability (degree of agreement between different raters) was calculated, as it is commonly performed in ASAG researches to better interpret the results. The agreement obtained 0.581 in Pearson's correlation coefficient, which was highly significant with a moderate correlation ($r = 0.581, p - value < 0.01$). This value is comparable to other literature datasets in other languages (more details in [Burrows et al. 2015]). The disagreement was removed to perform the experiments (only one grade was considered for each answer).

4. Proposed Method

The proposed approach to grade the answers is composed by four sets of features. Each group is described in the following subsections and then the preprocessing applied to each group is presented in a table in Subsection 4.5.

4.1. Bag-of-ngrams

Ngrams are one of the most common ways to model language in ASAG. It is based on the idea that the words' presence or absence can predict the desired output. The prob-

lem that arises from modeling text in this way is that no syntactical information is considered and sentences with opposite meaning using a negation word can be considered very similar. Despite the apparent naivety from using ngrams, it is still one of the most powerful predictors in ASAG context [Heilman and Madnani 2013, Magooda et al. 2016, Roy et al. 2016].

As ngrams works on the principle of presence or absence of text's pieces, it is a question-specific feature. Important words for a question are not important to another. Hence, each question has its own bag-of-ngrams sparse matrix of features, where each document is represented as a row and each ngram as a column. In this work, the simple term frequency was used, as term frequency-inverse document frequency did not perform as good.

Ngrams divides all the text in pieces considering n characters or words in sequence. For this work, the two types of ngrams were extracted: words and characters. To illustrate the difference, consider the sentence "*tem parede celular*". Word 2-grams of this sentence would be: ["*tem parede*", "*parede celular*"] and character 6-grams would return: ["*tem pa*", "*em par*", "*m pare*", " *pared*", ... , "*elular*"]. For both word and characters ngrams n ranged from 1 to 3. Only the top 300 (for words) and 350 (for characters) ngrams features were kept (concerning its importance, i.e. term frequency in the documents) totalizing 650 features in a sparse matrix.

4.2. Lexical similarity

This set of features is based on the lexical level similarity between the student's and reference's answers. This type of similarity is widely employed in ASAG research [Dzikovska et al. 2013, Sultan et al. 2016, Roy et al. 2016]. Four different groups of metrics are considered here to measure similarity, as grouped in [Vijaymeena and Kavitha 2016], described as follows:

1. **Token-based:** measures the similarity between two strings by considering the intersection of characters in both texts. Three different metrics were selected: Cosine, Overlap and Sorensen.
2. **Edit-based:** metrics of this type are based on counting the minimum number of operations performed to transform one string into the other. Levenshtein, Hamming and Jaro-Winkler were used.
3. **Sequence-based:** unlike token-based, here the order counts and similarity is based in sequences. One way is to measure the longest common substring between two given strings. The principle is that sentences with longest shared sequences are more likely to be similar. A variation of this idea is also employed in this work, using the RatcliffObershelp similarity.
4. **Compression-based:** is similar to edit-based but the similarity is extracted from the shortest computer program that can convert one string (in this case, represented as a bit vector) to another. The representative algorithm used was Normalized Compression Distance.

4.3. Semantic similarity

Semantic similarity in this work is measured using a semantic network (WordNet, as it is the most popular) to retrieve the semantic distance between words. In WordNet, words are

grouped in synsets, which are sets of cognitive synonyms, that represents some concept. The synsets are interlinked by their conceptual-semantic and lexical relations, providing means to measure semantic similarity.

There are few established algorithms that can compute word-to-word similarity in WordNet. They do so by walking through the links between synsets and measuring how close or distant they are, if they have hierarchical relationships, among other indicators. Six algorithms were used in this work: Leacock & Chodorow, Wu & Palmer, Lin, Resnik, Jiang & Conrath and Shortest Path.

In order to use word-to-word similarity for measuring answers similarity, an algorithm was implemented as proposed in [Mohler and Mihalcea 2009]. The idea is to compute the Cartesian product of the synsets from the words of the student and the reference answer, considering only open class words (nouns, verbs, adjectives and adverbs). Then, the six mentioned algorithms compute the similarities and add to a vector. Finally, the average and median of each metric is returned to be used as the features. This algorithm is applied for every pair (student answer, reference answer).

4.4. Other Features

The last group of features is composed of statistics extracted from each individual student answer and some ratio between them and the reference answers, described as follows.

Length Ratio: the length ratio between the student answer and the questions. Also, the maximum, minimum and mean of the ratio between the student answer and each reference answer. A large distance between the student and reference answer may indicate an incorrect answer. **Counts:** count per answer of: words, sentences, commas, unique words, negation words and each part-of-speech (POS) tag (in the universal POS tagset). Style of answers by the counts of their components may indicate better writing.

Word Length Average: the simple average of the length of words in the answer. Can indicate if answers with larger words turn in correct or incorrect grading. **Words per Sentence Average:** the size of each sentence. Another style writing feature to measure if shorter or larger sentences can lead to correct answers. **Concepts Match:** the match of keywords expected in correct student answers.

Another feature was extracted using **Latent Dirichlet Allocation**, a technique for automatically discover topics in the text. The probabilities of each answer belonging to one of the discovered topics are what is used as the feature.

4.5. Preprocessing

Five preprocessing techniques were considered in this work to improve the results. The application of each function to each feature set is presented in Table 2. They consisted in: **case normalization**, to not differentiate between upper and lowercase. **Non-alphanumeric characters removal**, as they do not add any value. **Accents removal**, to enhance matches between answers with and without accents. **Morphological reduction**, to make it easier to match words with only morphological differences. This can be accomplished with the use of lemmatization, an algorithm that reduces words to their root form. This is an important technique for Portuguese as it is a language with rich morphology. The last technique is **stopwords removal**, used to remove very common words so that when measuring similarity they are not taken into consideration.

Table 2. Preprocessing Application

X	Text Statistics	Bag-of-ngrams	LDA	Lexical Similarity	Semantic Similarity
Case normalization	No	Yes	Yes	No	Yes
Non-alphanumeric characters removal	No	Yes	Yes	No	Yes
Accents removal	No	Yes	Yes	No	No
Morphological reduction	No	Lemmatization	Lemmatization	No	Lemmatization
Stopwords removal	No	No	No	No	Yes

5. Experiments and Discussion

We experimented with the data by extracting the features as described in the previous section and combining them in different manners. All presented results were obtained using Extreme Gradient Boosting Classifier [Chen and Guestrin 2016] with default parameters and using 5-fold cross-validation. The Gradient Boosting algorithm was chosen because it is indicated by recent research to be one of the best in general [Zhang et al. 2017].

We performed experiments with automatic feature selection algorithms but due to their slowness, we opted for a simple threshold algorithm. It works by cutting off features using a threshold value on its importance, usually being the mean, but also the median or an empirical value.

First, our approaches are compared to a baseline classifier based on a simple class majority to predict samples. With the withdrawn of questions from ids 7 to 10, the data is reasonably balanced, as shown by the results achieved with the baseline classifier in Table 3, reporting accuracy scores for each question. Beyond the baseline, the table presents isolated results for each feature set of our approach, namely Ngrams, Text Statistics (TS), Lexical Similarity (LEX) and Semantic Similarity - WordNet (WN).

Table 3. Accuracy scores for 4-class classification

X	Baseline	Ngrams	TS	LEX	WN	Together	Voting	Hard S.	Soft S.
1	30,901%	**65,236%**	49,785%	51,931%	51,073%	61,373%	62,232%	59,657%	62,661%
2	34,034%	71,849%	64,706%	70,588%	57,143%	**75,210%**	69,748%	67,647%	67,227%
3	47,945%	68,950%	54,795%	58,447%	59,361%	67,580%	65,297%	**71,233%**	67,123%
4	44,253%	**82,184%**	61,494%	70,690%	65,517%	77,586%	74,138%	79,885%	76,437%
5	39,423%	**72,596%**	61,538%	60,096%	52,404%	66,346%	67,788%	64,423%	64,904%
6	30,952%	62,381%	52,857%	57,619%	50,000%	59,524%	61,905%	56,190%	**63,333%**
11	46,288%	65,066%	59,825%	63,319%	52,838%	61,135%	**67,686%**	65,066%	66,812%
12	42,157%	68,137%	62,255%	52,451%	56,863%	**68,627%**	64,216%	61,765%	67,647%
13	28,505%	53,738%	52,804%	55,607%	44,393%	56,075%	**58,879%**	55,140%	51,402%
14	43,534%	**69,397%**	54,741%	51,724%	50,000%	64,655%	59,052%	65,086%	62,931%
15	44,974%	77,249%	65,079%	68,254%	65,608%	74,603%	**78,307%**	68,254%	75,132%
Mean	**39,361%**	**68,798%**	**58,171%**	**60,066%**	**55,018%**	**66,610%**	**66,295%**	**64,941%**	**65,965%**

Next to them, results obtained from the combined features are presented, in four different manners. Firstly, the four feature sets are placed in the same feature vector, identified as *"Together"*. Secondly, we experimented with a Voting Classifier, that takes prediction of each feature set trained alone and computes the final prediction using the mode of the classes, identified in the table as *"Voting"*. The third approach was to use the class predictions of each feature set as input to a new classifier, using stacked generalization, identified as *"Hard Stacking (S.)"*. The fourth approach is a variation of the former, but using probabilities of each sample belonging to each label instead of the final predictions as input to a new classifier, identified in the table as *"Soft Stacking (S.)"*.

The best results obtained for each question are highlighted in Table 3. Is noticeable that the ngrams approach obtained the best scores in general, as it has the greatest values for four questions and the overall score. However, the other features also presented good results, as seen in the overall results, with text statistics, lexical similarity and semantic similarity obtaining 58%, 60% and 55% respectively, in comparison with almost 69% of the ngrams score.

Despite the fact the ngrams performed better than the other feature sets (the exception being in question 13, where lexical similarity is higher), it does not prevent the combined approaches from getting better results in some cases. The four last columns present the scores obtained by combining the features in different manners. The *Voting* approach obtained three best scores, *Together* got two and both *Stacking* won a single one each. Even though the combination of features has its strengths, ngrams performs better in general and the other sets are not higher enough in some cases to help improve the combined scores.

In Table 4 other metrics are reported for the average of the three best approaches from Table 3. Beyond these metrics, Table 4 also presents the confusion matrix obtained for all 11 questions, using the ngrams approach. In general, the confusion matrix shows that values closer to the main diagonal are higher whilst the distant ones are lower. This behavior shows the difficulty of the algorithm in splitting between adjacent classes.

A binary classification was also performed, turning grades 0 and 1 into 0 and grades 2 and 3 into 1. This is done in order to also test the ability of the system at differentiating mostly wrong than mostly correct answers. When performing this change, question 3 got very unbalanced and none of the approaches surpassed the majority baseline. Thus, it was consequently removed from the graphic in Figure 2 in order to improve visualization. There, we notice that the baseline model ranges from 50% to 70% and the other models from 65% to 90%. Again, ngrams usually perform better, but the other approaches are not too distant as they overlap among themselves right below the ngrams line.

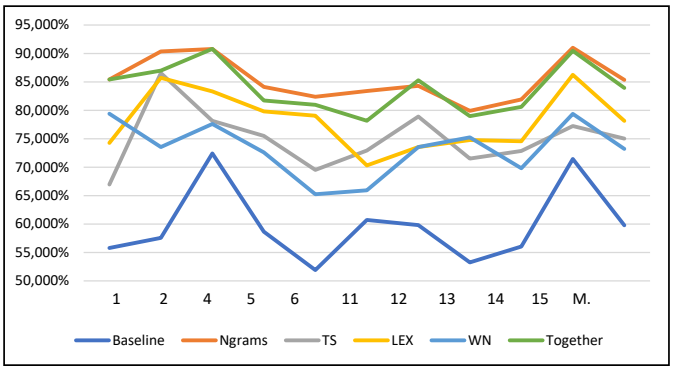

Figure 2. Binary Scores.

Table 4. Other scores and confusion matrix.

X	Ngrams	Together	Voting
Accuracy	0,68798	0,66155	0,66295
Precision	0,68300	0,65518	0,65232
Recall	0,68798	0,66155	0,66295
F1	0,68088	0,65365	0,64492
Kappa	0,54383	0,50378	0,49705
Confusion matrix (ngrams)			
603	92	54	32
91	375	110	28
46	103	369	60
16	25	84	262

6. Conclusions

In this work, we explored the Automatic Short Answer Grading research field. Through a literature review, we found some of the most relevant research on the field and briefly reported their characteristics. Then, we searched for works performed on Portuguese data and discovered that only a few researches have been performed, presenting some research opportunities.

In order to contribute to the development of the field, we presented a new Portuguese ASAG dataset, created with the participation of 245 students, 12 undergraduate students and 13 teachers, to be available for future researchers and comparison works. Furthermore, we reported our approach using four different sets of features, combined to seek for better results. We found that ngrams outperformed the other features groups in general, including their different combinations. Although the other three features sets did not perform as good, they also presented good results, being generally way ahead of the baseline and not too far from ngrams.

As future work, we want to explore more techniques addressed on other ASAG researches in order to improve results for our Portuguese dataset. We intend to do this because the ASAG task is modeled by different researches in a wide range of techniques and points of view, but still many of them are not widely experimented across different datasets. Another possibility is to explore more the specifics characteristics of the Portuguese language. Also, we expect to increase the number of dataset samples in order to apply deep learning techniques.

References

Ávila, R. L. F. and Soares, J. M. (2013). Uso de técnicas de pré-processamento textual e algoritmos de comparação como suporte à correção de questões dissertativas: experimentos, análises e contribuições. (Cbie):727–736. In Portuguese.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, pages 60–117.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., and Dang, H. T. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. *Seventh International Workshop on Semantic Evaluation*, pages 263–274.

Figueira, A. d. S., da Silva, A. G., de Melo, B. M., Lino, A. D. P., Lobato, F. R. L., and Favero, E. L. (2013). Módulo de Avaliação Automática de Questões Discursivas no Ambiente Virtual de Aprendizagem LabSQL. *Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5. In Portuguese.

Flores, E. M., Rigo, S. J., and Barbosa, J. L. V. (2014). Um modelo para avaliação automática de respostas textuais com uso de regras linguísticas. *Brazilian Symposium on Computers in Education (SBIE)*, 25(1):1153. In Portuguese.

Heilman, M. and Madnani, N. (2013). ETS: Domain Adaptation and Stacking for Short Answer Scoring. *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval)*, 2:275–279.

Higgins, D. e. a. (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *arXiv:1403.0801v2 [cs.CL]*.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., and Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, 53(2):215–233.

Magooda, A., Zahran, M. A., Rashwan, M., Raafat, H., and Fayek, M. B. (2016). Vector Based Techniques for Short Answer Grading. *International Florida Artificial Intelligence Research Society Conference Ahmed*, pages 238–243.

Mohler, M. and Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, pages 567–575.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Passero, G., Haendchen Filho, A., and Dazzi, R. (2016). Avaliação do uso de métodos baseados em lsa e wordnet para correção de questões discursivas. In *Brazilian Symposium on Computers in Education (SBIE)*, volume 27, page 1136. In Portuguese.

Ramachandran, L., Cheng, J., and Foltz, P. (2015). Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. *Workshop on Innovative Use of NLP for Building Educational Applications*, 10:97–106.

Riordan, B., Horbach, A., Cahill, A., Zesch, T., and Lee, C. M. (2017). Investigating neural architectures for short answer scoring. *$Bea17*, pages 159–168.

Roy, S., Bhatt, H. S., and Narahari, Y. (2016). An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading. 285:1622–1623.

Salton, G. D., Carniel, C. A., and Mello, B. A. D. (2013). Regras sintáticas livres de contexto na correção automática de Unidades de Leitura. pages 217–222. In Portuguese.

Santos, J. C. A., Ribeiro, T., Favero, E., and Queiroz, J. (2012). Aplicação de um método LSA na avaliação automática de respostas discursivas. *Revista Brasileira de Informática na Educação*, 0(0):10–19. In Portuguese.

Sultan, M. A., Salazar, C., and Sumner, T. (2016). Fast and Easy Short Answer Grading with High Accuracy. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075.

Vijaymeena, M. and Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Vilela, R. F., Valle, P. H. D., Muniz, R. J., Lima, W. A., Inocêncio, A. C. G., and Junior, P. A. P. (2012). SCATeDi : Sistema Inteligente para Avaliação de Desempenho Escolar em Avaliações Discursivas. *Workshop de Informática na Escola*, (1984):11. In Portuguese.

Zhang, C., Liu, C., Zhang, X., and Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128–150.

Zhang, Y., Shah, R., and Chi, M. (2016). Deep Learning + Student Modeling + Clustering: a Recipe for Effective Automatic Answer Grading. *Proceedings of the 9th International Conference on Educational Data Mining*, pages 562–567.