Automatic Grading of Portuguese Short Answers Using a Machine Learning Approach

Lucas B. Galhardi[1], **Rodrigo C. Thom de Souza**[2] **(Co-advisor),**
Jacques D. Brancher[1] **(Advisor)**

[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,jacques}@uel.br, thom@ufpr.br

Abstract. *Short answers are routinely used in learning environments for students' assessment. Despite its importance, teachers find the task of assessing discursive answers very time-consuming. Aiming at assisting in this problem, this work explores the Automatic Short Answer Grading (ASAG) field using a machine learning approach. The literature was reviewed and 44 papers using different techniques were analyzed considering many aspects. A Portuguese dataset was build with more than 7000 short answers. Different approaches were experimented and a final model was created with their combination. The model's effectiveness showed to be satisfactory, with kappa scores indicating moderate/substantial agreement between the model and human grading.*

1. Introduction

Assessments are routinely used in learning environments in order to estimate a percentage of the retained knowledge from students. Despite its importance, teachers usually find the task of assessing lots of discursive answers very time-consuming. The evaluation work is frequently done at home, compromising the teacher's life quality. Researches indicates that about 75% of some Brazilian teachers claims to frequently bring work home, like assessments of student's exams [Nascimento and Santos 2015]. This situation overloads teachers and reduces their time, that could be spent in other activities like class elaboration [Sakaguchi et al. 2015].

Teachers work's conditions and their own human subjectivity have a great impact on grading. Humans make mistakes and some reasons for that can be from fatigue, bias or the simple ordering of student's tests [Haley et al. 2007]. Moreover, with human grading, students may have to wait for a long time to receive feedback on their answers [Liu et al. 2016] and, when they finally get it, grades can be different from another classmate's, who has given a very similar answer [Passero et al. 2016, Santos et al. 2016].

These problems became more intense in tools like VLEs (Virtual Learning Environments) and MOOCs (Massive Open Online Courses), that have recently improved their popularity and are used by way more students than physical classes [ABED 2016]. Moreover, these environments can have assessment systems that can support teachers in evaluating many students. However, the assessment of written activities is frequently performed by humans, causing the previously exposed difficulties.

Computer-based assessment came to address these issues and improve other aspects of learning by automating the evaluation process. Some of the benefits of automatic assessments are: criteria is formalized [Williamson et al. 2012], can provide faster feedback to both teacher and student, can save teachers' time so they can use it to work better and allows teachers to easily follow the class performance [Santos et al. 2016]. Furthermore, automatic grading is becoming highly competitive with human grading, considering short answers [Butcher and Jordan 2010].

Evaluations are often composed of recall or recognition type of questions, which are in different levels of the learning depth. The recognition kind seeks to test the respondent's ability to organize or identify some specific information. As for the recall ones, 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].

Within the recall kind, there are questions concerning speaking, structured text (math and source code) or natural language questions. The natural language type can be classified in three groups: fill-the-gap, short answer and essay. Fill-the-gap expects responses to be only from one to a few words, with fixed openness and focus on words. Short answers varies from one sentence to one paragraph, the focus is on the content and it has closed openness. At last, essays can have from two paragraphs to several pages, focus is on the writing style and it has a more open scope [Burrows et al. 2015].

Considering this division, this work exclusively focus on short answers. In addition to the length, focus and openness, short answers must be written in some natural language and recalls to external knowledge outside the question statement. This research field is defined in [Burrows et al. 2015] as Automatic Short Answer Grading (ASAG). It consists in automatically assessing short natural language responses using computational methods. Several researches have been recently developed in the ASAG field. However, most of them uses English datasets, concerning the language of the questions and short answers. In its turn, ASAG research using Portuguese data is somewhat scarce. One of the reasons for this situation is the lack of public available Portuguese datasets, something that is not an issue for English research [Burrows et al. 2015].

1.1. Objectives, General Methodology and Outline

Considering the presented scenario, this work has as its main objective the exploration of the Portuguese ASAG field using a machine learning approach. The specific goals are the following: (1) Perform a systematic literature review of ASAG works that uses a machine learning approach; (2) Develop a web system to be used in ASAG context; (3) Put the system in operation to be used by several users and collect ASAG data from it. Then, publish the dataset on the web, to be publicly available; (4) Build an ASAG model to evaluate on the collected data and compare different approaches for the final model.

The presented goals are achieved through this work's general methodology, illustrated in Figure 1. This work is organized similarly to the presented goals. The following paragraphs presents an outline of this paper.

The first step was to perform a systematic literature review in order to get a overview of already existing works on the field (Section 2)

[Galhardi and Brancher 2018b]. With that done, a web system was modeled and developed in order to be operated by real world users [Galhardi and Brancher 2018a].

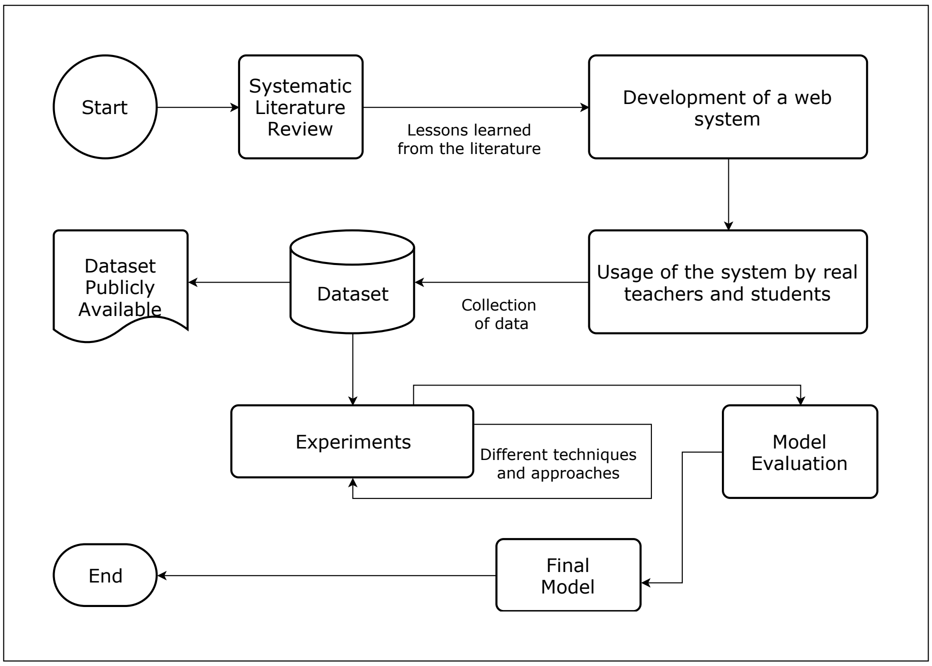
Start
Systematic Literature Review
Lessons learned from the literature
Development of a web system
Usage of the system by real teachers and students
Collection of data
Dataset
Dataset Publicly Available
Experiments
Different techniques and approaches
Model Evaluation
Final Model
End

Figure 1. General Methodology

From the system's usage from teachers and students, a dataset was collected and packaged in order to be made publicly available on the web (Section 3) [Galhardi et al. 2018].

From the dataset, several experiments were performed to test the performance of different approaches and techniques. Then, the different approaches were evaluated and compared. From them, a final model was created for grading new answers, composed of different groups of approaches (Section 6) [Galhardi et al. 2018]. Finally, Section 7 presents the final considerations for this work.

2. Systematic Literature Review (SLR)

First, the systematic review was planned, the protocol and research questions were defined and the inclusion, exclusion and quality criteria created. This SLR was performed as a lightweight SLR, where only one researcher assessed the papers in a short period of time (a few months). The final selection resulted in 44 papers and the six research questions were answered based on them.

The temporal distribution of studies revealed a stagnant state at first, but the scenario starts to change when six publicity available datasets were released in 2011, 2012 and 2013. Three of them came from different authors and three from two competitions, the ASAP's 2012 [1] and SemEval's 2013 [Dzikovska et al. 2013] ones. Those datasets opened the evaluation era in ASAG, where different techniques could now be directly compared.

[1] http://www.kaggle.com/c/asap-sas

Among the 44 papers, 28 different datasets were identified. A table gathering all datasets and their characteristics was created and eight aspects were analyzed. Datasets are usually in English, about science questions and from diversified age of respondents. The numbers of questions, answers and reference answers have large ranges of minimum and maximum values. The grading scale is usually from two to five classes and the responses length normally ranges from 7 to 63 words in average.

Many different NLP and preprocessing techniques are used among works and the most used ones are part-of-speech tagging, stemming, parsing, spelling correction and stopwords and other symbols removal. The features extracted to model answers can be grouped in four categories, the same as those studied in natural language processing: lexical, syntactical, semantic and discourse. These four aspects of natural languages are combined in order to model text as close as possible as the concepts they represent and the worthy grade of each answer. The selected features were used with several machine learning algorithms, with their frequency number of uses reported.

Finally, results achieved by the reviewed studies were grouped according to their correspondent dataset. Results were analyzed in terms of different agreement metrics between humans and proposed systems. Works that used private datasets were grouped and exposed the variety of metrics, classes and score values between different studies.

After performing the systematic review, the lack of works using Portuguese data was noticed. In order to fill this gap, we performed a simplified literature review aiming at finding some Portuguese related ASAG works. The main methodological differences are: (1) the primary source (only Google Scholar was used); (2) inclusion of papers written in Portuguese; (3) dropping the requirement for only machine learning approaches. After the search, seven papers were selected. Considering the review's results, the main difference from the other review is that, in the Portuguese case, datasets are usually way smaller; therefore, works do not employ a machine learning approach to grade the answers.

Both reviews were performed in 2017; hence results comprise only papers until 2016. An update of both reviews was performed in order to include the newest papers. For the Portuguese review, only two more works were found, very similar to those from previous years. For the original review, 15 new papers were found, reporting new findings and trends for ASAG.

A growing (and sophisticated) technique identified on new papers is the use of word embedding resources. However, despite its simplicity, features based on bag-of-words and ngrams are still very used in ASAG systems. Ngrams were also found to perform better than other features [Kohail and Biemann 2017], including word embedding based ones [Alvarado et al. 2018].

Another interesting observation in new papers is the confirmation of [Burrows et al. 2015] prediction of an "Evaluation Era": the majority (8/15) of works evaluates their system on one or more of the public datasets. However, despite the use of public datasets, results are often incomparable due to the use of different metrics among works. Also, authors do not search exhaustively for all recent work that uses the same dataset and evaluation metric as theirs, resulting in incomplete literature comparisons.

Other differences in newer ASAG research include: increase use of deep learning, data from others domains and sometimes coupled with other assessment items, new ideas for feature engineering and even adaptation of other field's techniques such as data augmentation.

3. Data Collection and Analysis

This section begins by introducing the *Auto-Avaliador* CIR web system, developed to be used in ASAG contexts. Then, the following subsections will go over the process of creating and analysing a new Portuguese ASAG dataset using the web system, the first to be publicity available and the data basis for this work (available at [2]).

3.1. The *Auto-Avaliador CIR* Web System

The *Auto-Avaliador CIR* system is a web environment developed to handle the dynamics involving questions and answers in a learning context. Using it, teachers can sign up and start creating tests and questions. In its turn, students have access to the tests and questions created by their teacher (and by others as well) and can answer and submit their responses. Student's answers will then be available for teachers to be assessed, providing feedback to students, that will see their grades in the system. For more details, screenshots and operation of the *Auto-Avaliador* CIR web system and its development process, please refer to [Galhardi and Brancher 2018a].

3.2. Data Collection

The system's first use came from five biology elementary school teachers from a Educational Professional Master class in the Pampa Federal University - Jaguarão/RS. Together, these teachers discussed and created one test with 15 questions in the *Auto-Avaliador CIR* system.

The subject matter addressed by the questions consists mainly of human body topics, mostly seen in the 8th grade of elementary school. Some examples are: *"Explique o mecanismo de inspiração e de expiração do ar no corpo humano"* (Explain the inspiration and expiration mechanism of the human body) and *"Quais são as diferenças entre veias e artérias?"* (What are the differences between veins and arteries?). 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 326 elementary school students (8th and 9th grades, about 12-14 years old) and to 333 high school students (10th to 12th grades, about 14-17 years). The application was made with the supervision of the student's teachers and each student had to come up with its own answers to the questions. Students were instructed to try their best, even if it involved guessing, in order to collect all sort of answers and grades.

In possession of the answers, 14 undergraduate biology students from the final year of college (from the same class) assessed the answers considering a predefined scale. Graders assigned one of four possible grades to each answer:

- **Zero:** when the answer is at least mostly wrong, out of scope or nonsense;
- **One:** if the answer has something correct but it is still mostly wrong or incomplete;

[2] https://www.kaggle.com/lucasbgalhardi/pt-asag-2018

- **Two:** if the answer is correct but has some wrong detail or missing important content;
- **Three:** if the answer is mostly correct, with the important points presented.

3.3. Data Analysis

The test containing 15 questions was applied to 659 students in total, so 9885 answers could be expected. However, students left some answers in blank. Additionally, in rare cases answers were completely equal, consisting in duplicated data, that was removed for all further analysis and experiments. The number of usable answers available was 7473, distributed between different grades as shown in Table 1.

Table 1. Labels' distribution

Q_ID	**0**	**1**	**2**	**3**	**Q_ID**	**0**	**1**	**2**	**3**
1	173	159	182	101	**9**	276	63	42	41
2	100	240	213	44	**10**	234	46	45	23
3	144	320	51	14	**11**	187	126	199	60
4	99	100	137	72	**12**	159	94	101	122
5	134	124	114	85	**13**	114	159	118	131
6	149	179	134	60	**14**	43	117	205	191
7	312	148	22	5	**15**	104	70	113	145
8	126	282	99	23	**Sum**	2354	2227	1775	1117

Most of the questions are reasonably balanced (10 out of 15, the imbalanced are: 3, 7, 8, 9 and 10), which is good for performing machine learning experiments (algorithms for imbalanced dataset were not tested). Each question has an average of 498 graded answers, ranging from 348 to 615.

3.4. Inter-rater Reliability

In the Introduction of this work, it was discussed that humans' subjectivity can have a great impact on grading. In this study, as in many others from ASAG, we measured how humans agrees between themselves regarding the grade to be assigned to an answer (the inter-rater reliability or agreement).

From the 15 questions, four (1, 9, 11 and 12) had all the answers graded by more than one rater. The metric chosen to measure the agreement is weighted Cohen's Kappa, as it is one of the most common and used metric for performing this kind of evaluation in ASAG [Liu et al. 2016] and in other domains and applications as well [Vanbelle 2016].

The weighted kappa statistic was created to account for different levels of disagreement [Cohen 1968]. So, if one answer is given 1 by one grader and 2 by another, the disagreement is not as weighty as if the grades were 1 and 3 (in the first case there is a 1-point disagreement and in the second case there is a 2-point disagreement).

Weights assigned to the kappa statistic are mainly linear or quadratic. The difference between them is that in the linear scheme the disagreement from 0 to 1 and from 1 to 2 is equally weighted, which is not the case for the quadratic approach, where the higher the two different scores, higher the *penalty*. It is recommended to report both weighting approaches when possible [Vanbelle 2016].

It is important to state that graders were all presented with the same criteria for assigning grades. Despite that, the disagreement between raters in the four analysed questions is very intense, as can be seen in Table 2.

Table 2. Disagreement between raters

Distance/ Kappa Score	Q1	Q9	Q11	Q12
Linear	0,4	0,43	0,39	0,37
Quadratic	0,57	0,54	0,52	0,5

Analysing the kappa scores according to the [Landis and Koch 1977] guidelines, graders have between fair to moderate agreement (0.2 - 0.4 it is considered "fair" and between 0.4 - 0.6 it is considered "moderate"). Interesting to observe that scores are similar among questions.

4. Methodological Procedures

The experiments started with the early definition of the methodology. Firstly, the evaluation metrics, preprocessing techniques, machine learning algorithms and implementation libraries were defined. Then, six different approaches were selected from the literature to model the answers. Each of these groups had its own experiments varying preprocessing techniques, machine learning algorithms and internal parameters (considering intrinsic characteristics of each group).

After testing the groups, their results were compared and discussed. Then, the best variant of each group was picked in order to create combinations between them aimed at improving the overall performance. Finally, the best combination model is compared to human performance. More details of all these procedures are provided in the remainder of this section and in Sections 5 and 6.

4.1. Evaluation Metrics

As related works in the literature uses a wide variety of metrics, all of them were considered. While confusion matrix and its derived metrics (Accuracy, Precision, Recall and F1) gives a good parameter for measuring performance, it lacks some issues. Firstly, these metrics can be deceiving if the data is imbalanced [Zhang et al. 2017], which is the case for some of the questions in this work. This could be remedy by a comparison with a dummy classifier (that assigns every sample with the most frequent class) [Dzikovska et al. 2012], but still it is harder to visualize the real performance.

The Cohen's Kappa correlation coefficient can better show the performance, way more independently of the data imbalance than accuracy. Also, it is a good measure for accounting for chance agreement [Liu et al. 2016, Moharreri et al. 2014]. Pearson's is also greatly used, but usually for cases when the measured variables are continuous. Spearman's rank also have its advantages in specific cases, but it is not too used in ASAG research [Galhardi and Brancher 2018b].

Considering the aforementioned aspects of different metrics and the ability of weighted Cohen's Kappa to measure ordinal variables (Subsection 3.4), we opted for two metrics: Cohen's Kappa (with linear and quadratic weights, as explained in Subsection

3.4) and Accuracy (for easy and quickly interpretability). However, accuracy is used only for informative purposes (only kappa is considered on comparisons: the average between the scores produced by both weighting schemes is used for direct comparison, namely from here on bk value). When not specifically stated in the following sections, scores are referring to the averaged bk score among all questions (specially in the graphic visualizations: the vertical axis reports this performance metric, which was used for general purpose comparisons).

4.2. Preprocessing

Five text preprocessing techniques were considered in this work when performing the experiments:

- **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 or stemming, algorithms that reduces words to their root or reduced form. This is an important technique for Portuguese as it is a language with rich morphology;
- **Stopwords removal:** used to remove very common words so that when measuring similarity they are not taken in consideration.

4.3. Machine Learning Algorithms

From the systematic review, the top-6 machine learning algorithms (in terms of use in works) are, respectively: Support Vector Machine (SVM[3]), Decision Tree (DT), Stacking, Logistic Regression, Random Forests (RF) and Gradient Boosting Machine (GBM). DT is greatly used mainly because of its capacity for interpretability, a desired ability in many applications. As for stacking, it is not a specific algorithm but a combination of others.

Moreover, SVM, DT, RF and GBM are in the top-5 performers in [Zhang et al. 2017] exhaustive machine learning algorithm's comparison. In order to choose between algorithms, we selected the intersection from the top-performers in [Zhang et al. 2017] with the top used from the review: Support Vector Machine, Random Forests and Gradient Boosting Machine. Decision Tree was left off as RF can be considered an improved version of DT. Stacking is used later for experiments that combines different approaches. All base models uses default hyperparameters settings[4]. All experiments are performed using 5-fold cross validation [Pulman and Sukkarieh 2005, Conort 2012, Madnani et al. 2013, Aldabe et al. 2015, Zbontar 2012]. The XGB (eXtreme Gradient Boosting) library is used in this work as the implementation of the Gradient Boosting Machine algorithm [Chen and Guestrin 2016].

5. Modeling Approaches

5.1. Ngrams

Ngrams is one of the most common ways to model language and a powerful predictor for ASAG [Heilman and Madnani 2013, Burrows et al. 2015, Roy et al. 2016]. It is based on

[3] With RBF (Radial Basis Function) Kernel

[4] Except for the n_estimators in Random Forest, changed to 100 to match with XGBoost and because its default value will also change to 100 in the next release version of the library.

the idea that words' presence or absence can predict the desired output. Using ngrams for ASAG modeling means that the learning algorithm will base the patterns' searches among the words used by the students. It will attempt to find which of the words (or sequences of characters) used by students correspond to correct answers and which do not.

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 (a document-term matrix), where each document (student answer) is represented as a row and each ngram as a column. Each cell contains a weight, that can be defined in different ways (but usually being the frequency).

5.2. Lexical Similarity

This set of features is based on the lexical level of language analysis. The features are extracted considering the similarity between students' and references' answers. This type of similarity is widely employed in ASAG research [Burrows et al. 2015].

In the lexical level, similarity is given by how related words appears to be, based on their constituents (letters). That means that two words that are very similar in shape (or even equals: homonyms) can be get a high similarity score, even if the meaning is not related (e.g. cell and cell phone). This problem is considered in this work and covered by semantic features in the next section.

Although only considering the lexical level, there are different groups of metrics that can be considered to measure similarity. Some of these metrics are grouped in [Vijaymeena and Kavitha 2016] survey in their *"String-Based Similarity"* section and used in this work, grouped by four different types:

1. **Token-based(3):** it measures similarity between two strings by considering the intersection of characters in both texts. Three different metrics were selected: Cosine, Overlap and Sorensen (Dice);
2. **Edit-based(3):** 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(2):** unlike token-based, here the order counts and similarity is based on 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(4):** it is similar to edit-based but 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. Four different variations were considered, depending on the compressor (*bwtrle*, *bz2*, *lzma* and *zlib*).

To extract the features, each student answer is compared against all the teachers' answers to the question. So, if there are three available teachers' answers for a question, the feature set for each student answer will consist of 36 features (3 reference answers × 12 metrics).

5.3. Semantic Similarity

As stated in the previous section, words that are similar in their shape can be not similar when considering their meaning. Also, words with very different shapes can have very close meaning. This property of natural languages brings a challenge to its correct processing by computers. Among other purposes, semantic networks were created to aid with this issue. The most representative and popular semantic network is WordNet [Miller 1995], a network where words are grouped in synsets, that are interlinked by their conceptual-semantic and lexical relationships, providing means to measure semantic similarity.

There are a 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 of these algorithms were considered for the experiments: Leacock & Chodorow, Wu & Palmer, Lin, Resnik, Jiang & Conrath and Shortest Path. These metrics are also used by many other ASAG works [Sakaguchi et al. 2015, Magooda et al. 2016, Roy et al. 2016]. The corpus used for statistical information required by the Resnik, Lin and Jiang & Conrath algorithms was the Mac-Morpho Brazilian Portuguese annotated corpus (words with their part-of-speech tags).

In order to use word-to-word similarity for measuring answers similarity, an algorithm was implemented as proposed in [Mohler and Mihalcea 2009], with the difference that here the median function was also experimented, in contrast with only the mean function of the original algorithm. Moreover, the use of both functions is also explored. The algorithm is applied to every $(studentAnswer, referenceAnswer)$ pair in the dataset.

5.4. Other Features

Another group of features that should be considered is Text Statistics. These features extracts some stats from the student answer. Also, there can be features extracted from some ratio between student and reference answers. There are many different types of text statistics used in ASAG literature. In this section, some of them are explored and used in the experiments. The list below presents the features used in this work, how many each group uses (there are 22 in total) and which works in the literature also uses it.

- **Length Ratio(4):** the length ratio between the student and the question statement. Also, the maximum, minimum and mean of the ratio between the student answer and the reference answers. A large distance between the student and reference answer may indicate an incorrect answer;
- **Counts(16):** count per answer of: characters, 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;
- **Average Word Length(1):** the simple average of the length of words in the answer. Can indicate if answers with larger words turns in correct or incorrect grading;
- **Words per Sentence (Average)(1):** the size of each sentence in terms of words. Another style writing feature to measure if shorter or larger sentences can lead to correct answers;

5.5. Word Embeddings

One of the greatest novelties in natural language processing in the last few years is word embeddings. The work of [Mikolov et al. 2013] introduced Word2Vec in 2013, a technique for representing words in vectors in an efficient manner. The authors presented two different models for accomplishing their goal: Skip-gram and CBow. From their work, several researches followed the embeddings path, leading to new and refined techniques.

In 2014, researches from Stanford University released GloVe (Global Vectors for Word Embeddings) [Pennington et al. 2014]. The main difference from the Word2Vec algorithm is that GloVe, beyond using context-based learning as Word2Vec, also uses global text statistics from the whole corpus by constructing a word co-occurrence matrix (like older methods such as Latent Semantic Analysis). This additional technique can improve the overall results, as demonstrated in their work [Pennington et al. 2014].

Following, in 2016 the Facebook Research team presented FastText [Bojanowski et al. 2016], a new extension to the Word2Vec algorithm. In Word2Vec, each word in the corpus is considered as an atomic entity, used for training the model. The novelty from FastText is that it treats words as a composition of character ngrams and hence, the vector for a specific word is determined from the sum of its character ngram's vectors. This representation difference can have a great impact depending on the data.

Training these word embeddings algorithms on a large corpus is a laborious activity. In order to help researchers to deal with natural language processing tasks in Portuguese, [Hartmann et al. 2017] trained word embeddings on a large Portuguese corpora, composed of 17 different corpus, totalizing 1.395.926.282 tokens. They used this data to train on the three aforementioned algorithms (Word2Vec, GloVe and FastText), making available different versions for each of them, concerning the model (Skip-gram or CBow) and the number of vector's dimensions.

6. Results and Discussion

The results achieved for each question by the six different approaches are presented in Table 3 (using the bk scores for comparison). The columns of the table are ordered from the best to the worst approach, from left to the right (considering the mean score in the last row). The table is colored in order to highlight discrepancies among questions. The colors are specially helpful at spotting questions where a specific approach was particularly bad or good. It also helps spotting questions that diverged from the more common pattern.

A first insight from Table 3 is that the techniques considering only words and their representation (NGRAMS and WEREP (Word Embeddings Representation)) got the higher scores. They also have a much larger dimensionality than the other feature's set (900's against dozens from the others). The only question that NGRAMS got a smaller score than WEREP was in question 7. For all the others, WEREP follows NGRAMS closely, but do not achieves the same performance.

A second attention caller is TXST (Text Statistics), holding the worst performance. This is expected as this group of features only accounts for simple text style statistics. Even though TXST got a bad performance compared to the others, it got reasonable scores for questions 2, 8 and 11, not that far away from the others. It even won from WNSIM in four of the questions.

Following, as intermediates, there are the three similarity approaches. The lexical similarity approach usually got higher scores than the other two. The difference between WE and WN (WordNet) similarity is smaller but WE usually performs better than WN. Interesting to notice that LEXSIM (Lexical Similarity) got a specially low score for question 9, losing by far from the other similarity approaches. Still regarding question 9, the WESIM (Word Embeddings Similarity) method got a score higher than WEREP and almost as the same as NGRAMS.

Table 3. Results from all modeling approaches side by side

Q_ID	NGRAMS	WEREP	LEXSIM	WESIM	WNSIM	TXST
1	0,612	0,586	0,560	0,447	0,543	0,328
2	0,645	0,636	0,672	0,646	0,557	0,573
3	0,575	0,475	0,481	0,439	0,415	0,321
4	0,775	0,728	0,667	0,654	0,647	0,448
5	0,685	0,638	0,659	0,591	0,530	0,519
6	0,487	0,452	0,446	0,387	0,394	0,242
7	0,364	0,389	0,401	0,292	0,377	0,286
8	0,495	0,442	0,445	0,439	0,357	0,399
9	0,484	0,461	0,323	0,476	0,425	0,296
10	0,820	0,714	0,769	0,576	0,658	0,449
11	0,517	0,501	0,478	0,479	0,390	0,417
12	0,592	0,572	0,502	0,446	0,511	0,347
13	0,362	0,326	0,317	0,232	0,216	0,176
14	0,639	0,599	0,582	0,595	0,461	0,475
15	0,808	0,794	0,764	0,741	0,761	0,600
Mean	0,591	0,554	0,538	0,496	0,483	0,392

Other discrepancies from the mean score sequence are from questions 2 and 7. For these questions, LEXSIM got the highest score, even greater than NGRAMS and WEREP. Another noticeable discrepancy is from question 9, in which WESIM wins from four approaches and gets real close to NGRAMS. It is by far the best question from WESIM.

Finally, concerning WNSIM, its score for question 7 was considerable good, even beating NGRAMS and losing only for WEREP and LEXSIM. However, as the second worst approach, question 7 is its only highlight. In fact, WNSIM performs so badly that it even loses for TXST in four questions (2, 8, 11 and 14) and it gets close of losing in another two (5 and 13).

In summary, the main insight from Table 3 is that despite of their general performance order, each approach has its advantages in specific questions. The reason behind this finding must be certainly explored by future researches. Preliminary and shallow analysis were performed and failed to give an answer to the question. The non-conformity among questions motivated an idea to increase the general performance by somehow combine all of the approaches, getting the best from each one (explored in the next subsection).

6.1. Combining different approaches

The six different approaches explained in previous sections were combined in order to seek for better performance. The groups were combined in different ways and the best variant was called “Soft-6”. The Soft-6 approach won from the best solo performer (ngrams) only for a very low margin. However, after a detailed analysis, it was concluded that it was better to use the Soft-6 approach instead of only ngrams.

6.2. Comparison with Human Grading

In this subsection, the agreement’s scores between human raters and between one of the human raters and the soft-6 model are compared. The scores are reported in Table 4 using linear and quadratic kappa for all questions where both scores are available (SHA: System-Human Agreement, HHA: Human-Human Agreement).

Table 4. HHA vs SHA agreement

Linear Kappa			Quadratic Kappa		
ID	**HHA**	**SHA**	**ID**	**HHA**	**SHA**
1	0,40	0,52	**1**	0,57	0,64
9	0,43	0,44	**9**	0,54	0,55
11	0,39	0,48	**11**	0,52	0,58
12	0,37	0,52	**12**	0,50	0,63
Mean	0,40	0,49	**Mean**	0,53	0,60

Considering the average between the four questions, both SHA and HHA and both linear and quadratic kappa scores are within the range of *moderate* agreement (as to [Landis and Koch 1977]’s guidelines, scores between 0.4 and 0.6). Except for question 9, that has very close scores between SHA and HHA, the other questions have a large difference between SHA and HHA. The results reported in Table 4 shows that the SHA scores are higher than the HHA scores. This means that there was more agreement between the soft-6 model and one of the raters than among human raters. This result is not particularly bad or good, but indicates that the model really learned from the scores assigned by human raters to the answers. Therefore, the model learned in such way that it disagrees less than two humans do between themselves. Also, it can indicates that the model performed really great or that some human raters misunderstood the assignment criteria (or even both cases mixed).

Regarding the literature results for the same kind of comparison, a considerable amount of them does not even report the HHA agreement. From those who does report it, it is not that common for SHA scores to be greater than HHA scores. However, it is not that unusual as well [Dzikovska et al. 2012, Moharreri et al. 2014, Ramachandran et al. 2015].

7. Final Considerations

The automatic grading of short answers is a valuable resource for the improvement of students’ evaluations. However, research using Portuguese data is scarce, specially considering works with a great amount of data, suitable for a machine learning approach. Considering this, this work explored the Automatic Short Answer Grading research field.

As the systematic review revealed a lack of research regarding Portuguese data, a new ASAG dataset was created. It was done by collecting data from the real world, counting with the participation of 659 students, 14 undergraduate students and 13 teachers. The ASAG dataset presented in this work, in the Portuguese language, is the first one, as far as we know, to be made publicly available. The dataset possesses a reasonable large amount of data, compared to other literature datasets. Its creation was intended to perform experiments using Portuguese tools, in order to test for the generality of another languages' techniques. Furthermore, it is made available so future researches can present and test new models against this dataset, reporting comparable results.

Results from the experiments showed that a simple approach can be as effective as a sophisticated one. However, a combination of all approaches presented the best general performance. The performance comparison between the best model and human grading showed that the agreement score is very similar, indicating that the model can be effectively used for grading and aiding teachers and educational institutions .

References

ABED (2016). *Censo EAD Brasil 2016 - Relatório Analítico de Aprendizagem a Distância no Brasil.* ABED.

Aldabe, I., Lacalle, O. L., Maritxalar, M., and Lopez-Gazpio, I. (2015). Supervised Hierarchical Classification for Student Answer Scoring. *arXiv preprint*.

Alvarado, J. G., Abdi Ghavidel, H., Zouaq, A., Jovanovic, J., and Mcdonald, J. (2018). A Comparison of Features for the Automatic Labeling of Student Answers to Open-ended Questions. *Edm*, pages 55–65.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*.

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.

Butcher, P. G. and Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, 55(2):489–499.

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.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Conort, X. (2012). Short Answer Scoring — Explanation of "Gxav" Solution. *ASAP '12 SAS Methodology Paper*, pages 1–22.

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.

Dzikovska, M. O., Nielsen, R. D., and Brew, C. (2012). Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 200–210.

Galhardi, L., Barbosa, C. R., de Souza, R. C. T., and Brancher, J. D. (2018). Portuguese automatic short answer grading. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1373.

Galhardi, L. and Brancher, J. D. (2018a). Auto-avaliador colaborativo e inteligente de respostas. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 142.

Galhardi, L. B. and Brancher, J. D. (2018b). Machine learning approach for automatic short answer grading: A systematic review. In *Ibero-American Conference on Artificial Intelligence*, pages 380–391. Springer.

Haley, D. T., Thomas, P., De Roeck, A., and Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: Using a computer to mark questions about HTML. *Conferences in Research and Practice in Information Technology Series*, 66:35–42.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. *arXiv preprint arXiv*.

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.

Kohail, S. and Biemann, C. (2017). Matching , Re-ranking and Scoring : Learning Textual Similarity by Incorporating Dependency Graph Alignment and Coverage Features. *18th International Conference on Computational Linguistics and Intelligent Text Processing. Budapest*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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.

Madnani, N., Burstein, J., Sabatini, J., and O'Reilly, T. (2013). Automated Scoring of a Summary-Writing Task Designed to Measure Reading Comprehension. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168.

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.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Moharreri, K., Ha, M., and Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(1):15.

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.

Nascimento, M. d. G. C. d. A. and Santos, J. V. (2015). Sessão Especial 05 - Políticas educacionais e currículo: interfaces na educação infantil e ensino fundamental. *37^a Reunião Nacional da ANPEd – 04 a 08 de outubro de 2015, UFSC – Florianópolis*.

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.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. *Proceedings of the second workshop on Building Educational Applications Using NLP - EdAppsNLP 05*, 1(June):9–16.

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.

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

Sakaguchi, K., Heilman, M., and Madnani, N. (2015). Effective Feature Integration for Automated Short Answer Scoring. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054.

Santos, J. C. A. d. et al. (2016). Avaliação automática de questões discursivas usando lsa. *Universidade Federal do Pará*.

Vanbelle, S. (2016). A New Interpretation of the Weighted Kappa Coefficients. *Psychometrika*, 81(2):399–410.

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.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Zbontar, J. (2012). Short Answer Scoring by Stacking. *ASAP '12 SAS Methodology Paper*, pages 1–7.

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.