Automatic Evaluation of Textual Cohesion in Essays

Aluizio Haendchen Filho[1], Filipe Sateles Porto de Lima[2], Hércules Antônio do Prado[2],
Edilson Ferneda[2], Adson Marques da Silva Esteves[1] and Rudimar L. S. Dazzi[1]
[1]Laboratory of Technological Innovation in Education (LITE), University of the Itajai Valley (UNIVALI), Itajai, Brazil
[2]Catholic University of Brasilia, Graduate Program in Governance, Technology and Innovation, Brasilia, Brazil

Keywords: Textual Cohesion, Automated Essay Grading, Machine Learning, Text Classification.

Abstract: Aiming to contribute to studies on the evaluation of textual cohesion in Brazilian Portuguese, this paper presents an approach based on machine learning for automated scoring of textual cohesion, according to the evaluation model adopted in Brazil. The purpose is to verify the mastery of skills and abilities of students who have completed high school. Based on features groups such as lexicon diversity, connectives, readability indexes and overlap of sentences and paragraphs, 91 features, based in TAACO (Tool for the Automatic Analysis of Cohesion), were adopted. Beyond features specifically related to textual cohesion, other were defined for capturing general aspects of the text. The efficiency of the classification model based on Support Vector Machines was measured. It was also demonstrated how normalization and class balancing techniques are essential to improve results using the small dataset available for this task.

1 INTRODUCTION

The national high school examination (known as ENEM) is an evaluation that happens annually in Brazil to verify the knowledge of the participants about various skills acquired during the school years. There are four exams consisting of multiple-choice tests, encompassing diverse contents, and a manuscript essay. The multiple-choice or objective questions are evaluated according to the response indicated, but the essay needs to be evaluated by at least two reviewers, which makes the process time-consuming and expensive. essays were evaluated in 2017 at an individual cost of U$ 4.96, totalling nearly U$ 32.45 million. This amount accounts for the structure, logistics and personnel needed to evaluate the national exam.

During the essay evaluation, two reviewers assign scores ranging from 0 to 200, in intervals of 40, for each of the five competencies that make up the evaluation model. Score 0 (zero) indicates that the author of the text does not demonstrate mastery over the competence in question. In contrast, score 200 indicates that the author demonstrates mastery over competence. If there is a difference of 100 points between the scores given by the two reviewers, the essay is analysed by a third one. If the discrepancy persists, a group of three reviewers (INEP, 2017) will evaluate the essay. The evaluated competencies are:

1. Domain of the standard norm of the Portuguese language.
2. Understanding the essay proposal.
3. Organization of information and analysis of text coherence.
4. Demonstration of knowledge of the language necessary for the argumentation.
5. Elaboration of a proposed solution to the problems addressed, respecting human rights, and considering the socio-cultural diversities.

A study of ENEM essays (Klein, 2009) shows that Competence 4 is one that poses a greatest challenge for students. For each competence, seven categories are established based on the scores. Two reviewers perform the corrections. Table 1 shows the proportion of scores given for each category, where category 1 refers to the lowest grade and category 7 refers to the highest grade for each competence.

Table 1: Proportion of scores by categories (Klein, 2009).

Compe tence	**Category of answer**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**Σ**
1	0.01	0.05	0.19	0.31	0.31	0.10	0.02	1.00
2	0.02	0.09	0.23	0.30	0.25	0.08	0.01	1.00
3	0.04	0.14	0.30	0.29	0.17	0.04	0.01	1.00
4	**0.05**	**0.15**	**0.30**	**0.28**	**0.16**	**0.05**	**0.01**	**1.00**
5	0.05	0.16	0.30	0.28	0.16	0.04	0.01	1.00

Filho, A., Porto de Lima, F., Antônio do Prado, H., Ferneda, E., Esteves, A. and Dazzi, R.
Automatic Evaluation of Textual Cohesion in Essays.
DOI: 10.5220/0011113600003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 634-640
ISBN: 978-989-758-569-2; ISSN: 2184-4992

Competence 4 is strongly linked to the author's ability to write a text in a cohesive, clear, and structured way. For this, the students use resources of textual cohesion. The difficulty around this competence is related to the difference between spoken and written language. While in a conversation the minimal grammatical structure is enough to convey a clearly message, in a text it is necessary to adopt a more formal and objective posture. As opposite the conversation, the text does not provide context signals easily perceivable by the reader senses (Shermis, Burstein, 2013). Therefore, those who fail to achieve high grading in this competence will have difficulty in articulating ideas cohesively through writing. Table 2 describes scores to be attributed for Competence 4.

Table 2: Descriptions of Competence 4 scores (INEP, 2017).

Score	Description
200 points	It articulates well the parts of the text and presents a diversified repertoire of cohesive resources.
160 points	It articulates the parts of the text with few inadequacies and presents a diversified repertoire of cohesive resources.
120 points	It articulates the parts of the text in a medium way, with inadequacies, and presents a little diversified repertoire of cohesive resources.
80 points	It articulates parts of the text insufficiently, with many inadequacies, and presents limited repertoire of cohesive resources.
40 points	It articulates the parts of the text in a precarious way.
0 points	Does not articulate information.

Systems for automatic grading of essays are built using several technologies and heuristics that allow evaluating with certain accuracy the quality of essays. Moreover, unlike human evaluators, these systems maintain consistency over the assigned scores, as they are not affected by subjective factors. They also help to reduce costs and enable faster feedback to the student-practicing essay (Tang, Suju, Narayanan 2009).

The main covered topics are: (i) a brief survey on the analysis of textual cohesion, (ii) the treatment of the corpus of essays extracted from the UOL and Escola Brazil sites; (iii) extraction and selection of specific features of textual cohesion; (iv) the use of Random Under Sampler for class balancing; (v) evaluation of the classification model based on the Support Vector Classifier.

2 BACKGROUND

Textual cohesion refers to the use of vocabulary and grammatical structures by means of connecting the ideas contained in a text. This connectivity property, also called contexture or texture by Textual Linguistics, is one of the aspects that promote good articulation and the logical-semantic structure of discourse (Koch, 1989).

The main mechanisms of textual cohesion are reference, substitution, ellipse, conjunction, and lexical cohesion. Each one is obtained by the proper use of cohesive links, elements that characterize a point of reference or connection in the text.

In order to simplify the analysis of textual cohesion, there are linguists which divide the mechanisms of cohesion into two groups: referential and sequential. In the first one, we considered the use of elements that retrieve or introduce a subject or something that is present in the text (endophoric reference), or outside the text (exophoric reference). The second encompasses the elements that give cadence and sequentially to the ideas presented in the text (Koch, 1989).

Take the following excerpt from the essay written by an ENEM participant from 2016 whose theme was "Pathways to combat religious intolerance in Brazil":

> ***Brás Cubas, the deceased-author of Machado de Assis,*** *says in his "Posthumous Memoirs" that (he) had no children and did not transmit to any creature the legacy of our misery. Perhaps today* ***he*** *perceived his decision to be correct: the attitude of* ***many Brazilians*** *towards religious intolerance is one of the most perverse aspects of a developing* ***society****. With this, there arises the problem of religious prejudice that persists is intrinsically linked to the reality of the country, whether by insufficiency of laws or by slow change of social mentality.*

The parts marked in bold highlight some references, such as the resumption of "Brás Cubas" in the apostrophe "the deceased-author of Machado de Assis" and the reference to "many Brazilians", which is an entity that is outside the text. The underlined portions indicate connectives as the discourse marker "with this". The idea pointed out in the previous sentence serves as a basis for the argumentation that follows. In addition, this passage presents an important property of the Portuguese Language: the reference by ellipse, indicated by the occurrence of "(he)" that was not originally included in the text. The ellipse consists in the omission of the subject before verbs, when it is possible to infer to whom or to what the action refers.

When analysing textual cohesion, it is necessary to verify, for example: (i) whether the references agree on number and gender with those referenced; (ii) whether the meaning of the connectives are in accordance with the context in which they are inserted; (iii) if the author avoided the repetition of terms; and (iv) whether ideas are connected logical and sequentially. That is, the analysis of textual cohesion has a very dynamic nature since it reflects flexibility of language use. However, a fact relevant to this analysis is that all information about textual cohesion exists in the text itself. Unlike textual coherence, which depends on the reader's knowledge of the world, cohesion is a strictly lexical-grammatical phenomenon (Halliday, 1976).

Assuming that cohesion is fully contained in the text, tools such as coh-metrix[1] and TAACO[2] have been constructed to identify and measure the parts of text that constitute the phenomenon of textual cohesion. Both compute similar metrics that comprise several dimensions of cohesion: *(i)* local cohesion, which exists between sentences; *(ii)* global cohesion, which exists in relation to the entire text; and *(iii)* lexical cohesion, which emerges from the use of the lexicon. These metrics are used to measure the quality of writing, the readability of the text, to verify the variation of the speech among other applications (Graesser et al 2014).

The process of obtaining these metrics is based on the use of natural language processing techniques, such as tagging and morphological normalization, textual segmentation and coreference analysis. The outcome of this process does not necessarily indicate the quality of use of the cohesion devices but provides information that enables further analysis.

3 TEXTUAL COHESION EVALUATION

The proposed approach was developed using the Feature-Based Engineering Method by means of the following steps: (i) organization of the corpus; (ii) extraction and normalization of features; (iii) class balancing, and (iv) classification. These steps are described as follows.

3.1 The Corpus of Essays

The essays used to construct the corpus that enabled our experiments were obtained through a crawling process of essays datasets from the UOL and Brazil School[3] portal.

Both portals have similar processes for the accumulation of essays: monthly a theme is proposed and interested students submit their textual productions for evaluation. Part of the essays evaluated are then made available on the portal along with the respective corrections, scores and comments of the reviewers. For each essay, a score between 0 and 2 is assigned, varying in steps of 0.5 for the 5 competences corresponding to the ENEM evaluation model.

To avoid possible noise in the automatic classification process, we perform the following processing steps:

1. Removal of special characters, numbers and dates.
2. Transformation of all text to lowercase.
3. Application of morphological markers (POS tagging) using the nlpnet library.
4. Inflection of the tokens by means of stemming using the NLTK library and the RSLPS algorithm, specific for the Portuguese language.
5. Segmentation (tokenization) by words, sentences, and paragraphs.

In addition to these steps, only the essays with more than fifty characters and whose scores available in all competencies were considered. Table 3 presents the general characteristics of the corpus after preprocessing.

Table 3: General metrics on the essay's corpus.

Metric	Value
Number of essays	8584
Number of average words per essay	269.84

3.2 Features Extraction and Normalization

Similarly, to Júnior, Spalenza and Oliveira (2017), each essay was represented as a feature vector. In

[1] Coh-metrix is a computational tool that produces indexes on discourse. It was developed by Arthur Graesser and Danielle McNamara. Tool documentation is available at http://tea.cohmetrix.com/.

[2] TAACO, as well as coh-metrix, produces measures on the linguistic characteristics of the text, but is more focused on textual cohesion metrics. The tool is available in http://www.kristopherkyle.com /taaco.html.

[3] Both extractions are available at https://github.com/ gpassero/uol-redacoes-xml.

total, 91 metrics of textual cohesion were calculated, based on those established by the TAACO system, with the appropriate adaptations to the Portuguese language. The features comprise several dimensions of lexical diversity, readability indexes, counting of connectives and measures of word overlap between sentences and between paragraphs.

Table 4: Characteristics of textual cohesion extracted from the corpus.

Type	Description
Lexicon Diversity and connectives (16 metrics)	Metrics that indicate how varied is the use of the lexicon in textual production. They were calculated from the token-type ratio and encompassed content words, functional words, verbs, adjectives, n-grams, pronouns, among others. In addition, we also calculated the incidence between connectives and some sentences. These features are directly related to cohesion sequence.
Readability Indexes (4 metrics)	The readability indexes measure how easy it is to read the text in relation to lexical diversity, word complexity, sentence size, among other factors. MTDL, HDD, ARI, and CLI were calculated.
Overlap of sentences and paragraphs (71 metrics)	To identify the referential cohesion relations in the text, several overlapping indexes were calculated. For example, overlapping names and pronouns between adjacent sentences and paragraphs, overlapping of adjectives, verbs, adverbs, words of content, among others.

As mentioned in Géron (2017), the standardization of the statistical distribution of features directly influence the quality of the machine learning model because it reduces the negative effect that outliers may cause during the training process. Then, to ensure the good performance of the model, z-score standardization was applied.

3.3 Classes Balancing

It is clear that the unbalanced number of essays per grade in Competency 4 (see Table 5) can negatively affect the classifier efficiency. To solve this problem, an approach based on the SMOTE (Synthetic Minority Oversampling Technique) algorithm was adopted. This algorithm searches the neighbours closest to the samples that have low representation in relation to the other classes of the dataset. From these neighbours, which have characteristics similar to the sample in question, the algorithm calculates a new sample to reinforce the number of examples in each class Chawla (2002). In this way, the set of examples available for classifier training was reinforced (Table 6), minimizing the impact that the class imbalance would cause in the classifier results.

Table 5: Number of essays per score in Competence 4 in the training set.

Score	Number of essays
0	**1106**
50	158
100	**4786**
150	288
200	529
Total	6867

Table 6: Number of essays per score in Competence 4 in the training set after class balancing.

Score	Number of essays
0	4786
50	4785
100	4786
150	4786
200	4785
Total	23.928

3.4 Classification

Training of the learning model was done using the stratified cross-validation method with k = 10, that is, the already normalized, balanced, and selected characteristics matrix were divided into ten equal parts, with each part containing examples of all classes. In this way, there were ten training iterations, so that in each iteration nine parts were used to train and one part to test.

As described by Júnior, Spalenza and Oliveira (2017), the problem of evaluating textual cohesion was treated as a classification problem where each essay receives a score between 5 possible scores. The strategy employed was to train a classification model. The learning algorithm used was the Support Vector Machine with linear core and C = 7 penalty of one-against-all type, that is, for each class a binary classifier was trained. This algorithm was chosen to generalize well in large dimensions in a consistent and robust way (Joachims, 2005).

4 DISCUSSION AND RESULTS

To avoid an overfitting situation, which occurs when the model fits the training data but does not generalize well to unknown instances, the test step was performed with a separate data set. It was generated early in the model building process and has representation in all possible scores that can be attributed to the essay. In this case, we decided that the test set would be equivalent to 20% of the essays available in the corpus.

To measure the performance of the learning model, the classical precision and recall metrics were calculated Júnior, Spalenza and Oliveira (2017) and Geron (2017), as presented in Table 7.

Table 7: Number of essays per score in Competence 4 in test set.

Score	Precision	Recall	Number of essays
0	0.28	0.47	276
50	0.04	0.25	40
100	**0.74**	**0.21**	**1197**
150	0.07	0.32	72
200	0.12	0.33	132
Average / Total	0.58	0.26	1717

We observed that even after applying balancing classes, the model obtained low precision and recall for classes with little representation, such as the cases of the 50 and 150 scores. On the other hand, the result provided by the model shows more adequate than the unbalanced form. Without this balance between classes, the model would present a high precision and general recall but based only on the dominant class. Another important observation is that due to SMOTE balancing (Section 3.3), the recall for dominant classes decreases to maintain balance with the other classes.

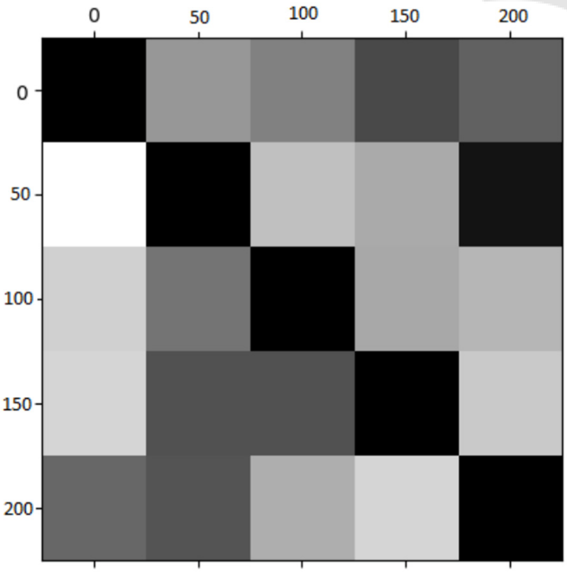

Figure 1: Matrix of confusion for the classifier that evaluates textual cohesion.

To better understand the errors made by the classifier, a confusion matrix was generated (Figure 1). It indicates, in the clearest parts, which classification is wrong. In the darkest parts it shows the correct classification for each score.

Although the apparent bad results shown in this confusion matrix, it is possible to argue in favour of the proposed approach by considering the number of essays concentrated in 100 and 150 grades. From 6,867 essays, 4,786 (near 70%) is concentrated in this region. Additionally, the bad results refer to the minority classes.

5 RELATED STUDIES

The automatic evaluation of essays characterizes a multidisciplinary area of study that encompasses linguistics, education, and computing. In this context, several works are carried out with the aim of developing new techniques that facilitate the application of these methods in production scales. Pioneers in this area, Page and Paulus (1968) proposed a system based on statistical methods that associate the writing style with the final attributed score of textual production. However, this analysis was done only by shallow features and disregarded the content of the text.

In order to develop systems that go beyond a superficial analysis and that are able to provide feedback to the student, new methods based on machine learning and natural language processing have been developed, now considering features such as grammatical assertiveness, adherence to the proposed theme, checking of facts Tang, Suju Narayanan, (2009). Thus, the scores attributed by these systems are based on a model closer to that used by human evaluators.

In Brazil, we find some approaches to the evaluation of automated essay scoring as a whole, evaluating production without turning to specific points such as grammar, syntax or theme. Among works that are in this category, it can be cited Passero et al. (2016) and Avila & Soares (2013). These works start from a strategy based on textual and semantic similarity, respectively, between the text written by the student and texts references that contain answers considered ideal. These methods are mainly used for automatic short answer grading and are based on metrics such as Levenshtein's distance and semantic similarity models such as Latent Semantic Analysis (LSA) or WordNet.

In a more focused way, some work on grading of ENEM essays treats specific competences as in Nau et al. (2017), where language deviations, one of the criteria evaluated in Competence 1 of the ENEM evaluation model, are detected based on a set of predetermined linguistic rules. This system provides a valuable input for more complete approaches related to Competence 1 evaluation. Another work also based on the ENEM model was developed by Passero, Haendchen Filho, Dazzi (2016) where Competence 2 regarding the deviation of the proposed theme is treated and provides excellent results.

Júnior, Spalenza and Oliveira (2017) presented a framework based on machine learning and natural language for the evaluation of Competence 1 of ENEM. The authors establish a set of features specific to Competence 1, as well as various ways of refining these characteristics in order to generate a machine learning model that achieves good results in the essay corpus of the Brazil School.

On the evaluation of textual coherence, some works propose ways of measuring this characteristic of the text. The TAACO system (Crossley, Kyle, & McNamara, 2016) and the coh-metrix (Graesser, McNamara, McCarthy, 2014) are reference tools in this context. In addition, extensive research was carried out on more specific points of textual coherence such as the analysis of co-referencing and the use of cohesive links for the summarization of texts.

6 CONCLUSIONS AND FUTURE WORKS

The automated analysis of textual cohesion presents several challenges, mainly related to the processing of features suitable for its characterization. The shortage of data and tools for the Portuguese language also worsen the situation, and more work on developing and improving NLP tools in Portuguese is needed.

One of the contributions of this work is the corpus of ENEM-based essays that is made available ready to use (download from <blind review>). This is relevant for research in Portuguese, beyond the usual English. Furthermore, the work introduces a set of textual cohesion features adapted to Portuguese. The adaptation had considered the linguistic differences at the morphological and syntactic levels between English and Portuguese. These publicly available features can be explored in other models of machine learning for the problem approached.

Regarding accuracy, the confusion matrix shows that the best results were obtained in the dominant classes, those that hold more than 80% of the occurrences in the scores. On the other hand, there is a need for methods capable of obtaining more precision in the attribution of scores close to the extremes.

The study also showed that gains in accuracy can be obtained for true positives by applying balancing techniques.

As future work, it is suggested: (i) to expand and improve the quality of the essays corpus; (ii) to evaluate other learning models based on neural networks and deep learning; (iii) to explore the lexical cohesion part and (iv) to compare the results here presented for Portuguese with those in other languages, say, English, for example.

REFERENCES

Avila, R. L. F.; 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. SBIE.

Chawla, N. V. et al. (2002) SMOTE: Synthetic Minority oversampling technique. Journal of Artificial Intelligence Research 16:321–357. https://www.jair.org/media/953/live-953-2037-jair.pdf

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016) The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods 48(4).

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2017). Redação no ENEM 2017: Cartilha do participante.. http://download.inep.gov.br/educacao_basica/enem/guia_participante/2017/manual_de_redacao_do_enem_2017.pdf

Géron, A. (2017) Hand-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly.

Graesser, A. C.; McNamara, D. S; McCarthy, P. M. (2014) Automated Evaluation of Text and discourse with Coh-metrix. Cambridge University Press.

Halliday, M. A. K; Hasan, Ruqaiya. (1976) Cohesion in English. Routledge.

Joachims, T. (2005) Text categorization with support vector machines: learning with many relevant features. European Conference of Machine Learning.

Júnior, C. R. C. A; Spalenza, M. A.; Oliveira, E. (2017) Proposta de um sistema de avaliação automática de redações do ENEM utilizando técnicas de aprendizagem de máquina e processamento de linguagem natural. Computer on the Beach. https://siaiap32.univali.br/seer/index.php/acotb/article/view/10592

Klein, R.; Fontanive, N. (2009) Uma nova maneira de avaliar as competências escritoras na Redação do ENEM. Ensaio: Avaliação e Políticas Públicas em

Educação. http://www.redalyc.org/articulo.oa?id=399537967002

Koch, I. V. (1989) A coesão textual. Brasil: Editora Contexto, 1989.

Nau, J. et al. (2017) Uma ferramenta para identificar desvios de linguagem na língua portuguesa. Proceedings of Symposium in Information and Human Language Technology (STIL). http://www.aclweb.org/anthology/W17-6601.

Page, E. B.; Paulus, D. H. (1968) The Analysis of Essays by Computer. Final Report. Connecticut Univ., Storrs. Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research, Bureau No-BR-6-1318. Pub Date Apr 1968.

Passero, G. et al. (2017) Off-Topic Essay Detection: A Systematic Review. CBIE. http://br-ie.org/pub/index.php/sbie/article/viewFile/7534/5330

Passero, G.; Haendchen Filho, A.; Dazzi, R. L. S. (2016) Avaliação do uso de métodos baseados em LSA e WordNet para Correção de Questões discursiva. SBIE. http://brie.org/pub/index.php/sbie/article/viewFile/6799/4684

Pinker, S. (1994) The language instinct. William Morrow and Company. 1994.

Shermis, M.; Burstein, J. (2013) Handbook of Automated Essay Evaluation: Current applications and new directions. Routledge.

Tang, L; Suju, R; Narayanan, V. K. (2009) Large Scale multi-label classification via metabeler. Proceedings of the 18th International Conference on World Wide Web.