Automatic Evaluation of Textual Cohesion in Essays

Filipe Sateles Lima[1], Aluizio Haendchen Filho[2], Hércules Antonio do Prado[1], Edilson Ferneda[1]

[1] Catholic University of Brasília, Brazil
[2] UNIFEBE - University Center of Brusque, Brazil

filipe.sateles@outlook.com, aluizio.h.filho@gmail.com, Hercules@ucb.br, eferneda@pos.ucb.br

Abstract. This paper presents an approach based on machine learning for automated essay grading of textual cohesion according to the evaluation model adopted in Brazil to verify the mastery of skills and abilities of students who have completed high school. From specific textual cohesion features and features that are able to capture general aspects of the text, we trained and measured the efficiency of a classification model based on support vector machines. Furthermore, we demonstrate how normalization and class balancing techniques are essential to improve our results using the small dataset available for this task.

Keywords: Textual cohesion. Automated essay grading. Machine learning. Text classification.

1. Introduction

The national high school examination (known as ENEM) is an evaluation that happens annually in Brazil in order 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. According to a survey carried out by the G1[1] portal, 6.54 million essays were evaluated in 2015 and each cost U$ 4.96, totaling approximately 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.

[1] Data available in <https://g1.globo.com/educacao/enem/2016/noticia/corretores-de-redacao-do-enem-avaliam-em-media-74-redacoes-por-dia.ghtml>. Last accessed: 10/06/2017.

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 analyzed by a third one. If the discrepancy persists, a group of three reviewers [4] will evaluate the essay. The competencies evaluated 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 the 2008 ENEM essays [10] 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.

Competence	**Category of answer**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**Σ**
1	0.010	0.054	0.189	0.309	0.314	0.105	0.019	1.000
2	0.024	0.091	0.229	0.305	0.254	0.083	0.015	1.000
3	0.045	0.143	0.301	0.287	0.172	0.045	0.007	1.000
4	**0.052**	**0.153**	**0.298**	**0.280**	**0.165**	**0.046**	**0.007**	**1.000**
5	0.050	0.156	0.300	0.284	0.163	0.042	0.006	1.000

Table 1. Proportion of scores by categories [10]

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 [15]. Therefore, those who fail to achieve high grading in this competence will have difficulty in articulating ideas cohesively through writing. Table 2 describes scores that can be attributed to Competence 4.

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 [16].

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.

Table 2. Descriptions of Competence 4 scores [4]

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 [11].

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 sequentiality to the ideas presented in the text [11].

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 analyzing 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 logically 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 [7].

Assuming that cohesion is fully contained in the text, tools such as coh-metrix[2] and TAACO[3] 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, in order to verify the variation of the speech among other applications [6].

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 correference 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. Proposed Approach

The proposed approach was developed 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

[2] 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/>. Last access on 01/01/2018.

[3] 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 and documentation is available in <http://www.kristopherkyle.com/taaco.html>. Last access on 01/02/2018.

School[4] 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.

In order 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;
6. The scores attributed to each competence were normalized according to the ENEM model (0 to 200 varying in steps of 50). This procedure was necessary to correct the scale of some notes that were assigned as 5, 10, 20 instead of 0.5, 1.0 and 2.0, for example. In this way, all the notes were on the same scale.

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.

Metric	Value
Nº of essas	8584
Nº average words per essay	269.84

Table 3. General metrics on the essays corpus.

3.2. Features extraction and normalization

Similarly to Júnior, Spalenza and Oliveira [9], each essay was represented as a feature vector. In 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[5].

[4] Both extractions were carried out by Guilherme Passero and are available at: https://github.com/gpassero/uol-redacoes-xml, Last accessed: 11/01/2017.

[5] The complete table with the textual cohesion characteristics extracted to train the model presented in this paper, as well as the code used, is available in https://github.com/sateles/nexo.

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)	In order 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.

Table 4. Characteristics of textual cohesion extracted from the corpus

As mentioned in Géron [5], 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, where *X* represents the characteristic matrix, *υa* represents the average of each line of *X*, and σ the standard deviation of each line of *X*.

$$Zscore = \frac{X - \upsilon}{\sigma}$$

3.3. Classes Balancing

We observed that the corpus used for the experiments had an unbalanced amount of essays per grade in Competence 4 (see Table 5), and that this could negatively affect the efficiency of the classifier. To solve this problem, our approach is based on the SMOTE (Synthetic Minority Oversampling Technique) algorithm. This algorithm searches the neighbors closest to the samples that have low representation in relation to the other classes of the dataset. From these neighbors, which have characteristics similar to the sample in question, the algorithm calculates a new sample to reinforce the number of examples in each class [2]. 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.

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

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

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

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

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 iteraction nine parts were used to train and one part to test.

As descrtibed by Júnior, Spalenza and Oliveira [9], 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 [8].

4. Discussion and Results

In order 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 [9,5], as presented in Table 7.

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

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

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 as a whole 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 in order to maintain balance with the other classes.

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.

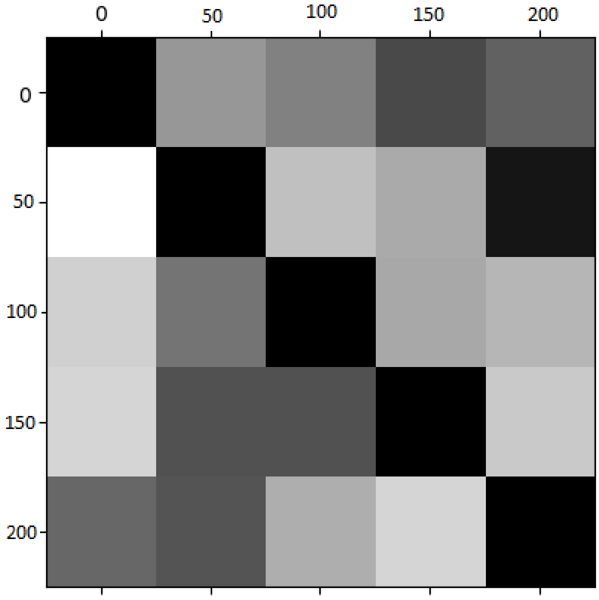

Figure 1. Matrix of confusion for the classifier that evaluates textual cohesion.

5. Related Works

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 (1998) 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 (Yigal and Burstein, 2005,16]. 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. [14] and Avila [1]. 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 Analisys (LSA) or WordNet.

In a more focused way, some work on grading of ENEM essays treats specific competences as in Nau et al. [12], 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 et al. [13], where Competence 2 regarding the deviation of the proposed theme is treated and provides excellent results.

Júnior et al. [9] 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 [3] and the coh-metrix [6] 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. Final Considerations and Future Works

The automatic 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 aggravate the situation, and more work on developing and improving NLP tools in Portuguese is needed.

Although the results obtained here are not ideal for a real production environment, this work introduces a set of textual cohesion features adapted to Portuguese. These features can be explored in other models of machine learning in order to evaluate textual cohesion.

As future work, we plan: *(i)* to expand and improve the quality of the essays corpus; *(ii)* evaluate other learning models based on neural networks and deep learning; *(iii)* improve the set of features with more effective reference analysis techniques; and *(iv)* further explore the lexical cohesion part.

References

[1] Avila, R. L. F.; Soares, J. M. Uso de técnicas de pré-processamento textual e algoritmos de comparação como suporte à correção de questões dissertativas: experimentos, análises e contribuições. SBIE, 2013.

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

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

[4] DAEB, Diretoria de avaliação da educação básica. Redação no ENEM 2017: Cartilha do participante. INEP, 2017. <http://download.inep.gov.br/educacao_basica/enem/guia_participante/2017/manual_de_redacao_do_enem_2017.pdf>

[5] Géron, A. Hand-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly. 2017.

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

[7] Halliday, M. A. K; Hasan, Ruqaiya. Cohesion in English. Routledge, 1976.

[8] Joachims, T. Text categorization with support vector machines: learning with many relevant features. European conference of machine learning, 2005.

[9] Júnior, C. R. C. A; Spalenza, M. A.; Oliveira, E.. 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, 2017. <https://siaiap32.univali.br/seer/index.php/acotb/article/view/10592>.

[10] Klein, R.; Fontanive, N. Uma nova maneira de avaliar as competências escritoras na Redação do ENEM. Ensaio: Avaliação e Políticas Públicas em Educação [en linea], 2009. <http://www.redalyc.org/articulo.oa?id=399537967002>

[11] Koch, I. V. A coesão textual. Brasil: Editora Contexto, 1989.

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

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

[14] Passero, G.; Haendchen Filho, A.; Dazzi, R. L. S. Avaliação do uso de métodos baseados em LSA e WordNet para Correção de Questões discursiva. SBIE, 2016. <http://br-ie.org/pub/index.php/sbie/article/viewFile/6799/4684>.

[15] Pinker, S. The language instinct. William Morrow and Company. 1994.

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

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