Exploring the Automated Scoring of Narrative Essays in Brazilian Portuguese using Transformer Models

Eugénio Ribeiro[1] and **Nuno Mamede**[1,2] and **Jorge Baptista**[1,3]
[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract

The automated scoring of narrative essays written by students according to different competences can assist teachers in their evaluation process and help them to focus on specific areas of writing that require improvement among their students. In this paper, we explore the fine-tuning of Portuguese foundation models to automatically score student essays according to four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion. The results of our experiments show that the agreement between these models and human graders varies between fair and substantial. Thus, although they can provide cues for essay scoring, significant research is still required towards their improvement, especially for the more complex competences.

1 Introduction

Automated Essay Scoring (AES) has garnered significant attention due to its potential to revolutionize the assessment of written language, particularly in educational settings. The PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays addresses this problem in the context of the Brazilian basic education system by focusing on scoring essays according to four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion.

In this paper, which describes our approach to the competition, we explore how Transformer-based foundation models for Portuguese perform when fine-tuned for scoring essays in terms of each of the target competences. With this study, we aim to assess whether these models are sufficiently robust for the task and can help teachers in their evaluation process or whether additional information is required to make them useful.

2 Related Work

AES has evolved considerably since its inception (Ifenthaler, 2022). The work by Haswell (2006) provides comprehensive insights into the development and history of AES. Recent reviews (e.g., Uto, 2021; Ramesh and Sanampudi, 2022; Vijaya Shetty et al., 2022), offer up-to-date perspectives on the state-of-the-art techniques and challenges in AES. Ethical considerations surrounding AES implementation, including economic pressures and validity concerns, have been extensively discussed (Jones, 2006; McAllister and White, 2006; Hannah et al., 2023). Furthermore, studies have explored the quality assessment of AES systems, aiming to maximize agreement between human and machine evaluations (Chen and He, 2013). Recent advancements in deep learning have propelled AES, with Transformer models and multimodal machine learning approaches gaining traction (Zhu and Sun, 2020; Kumar and Boulanger, 2021; Ludwig et al., 2021). Evaluation campaigns on AES (Mathias and Bhattacharyya, 2020) signal advancements in the area and can potentially enhance the efficiency and effectiveness of essay assessment processes.

In Brazilian Portuguese, research on AES has mainly focused on automatically grading the Exame Nacional do Ensino Médio (ENEM), which serves as an admission test for most universities in Brazil. Recent advances on this subject were mainly based on the development of the Essay-BR corpus (Marinho et al., 2022) and the fine-tuning of foundation models (Matsuoka, 2023). However, this problem had already been explored using both frequency-based (Bazelato and Amorim, 2013) and manually engineered features (Amorim and Veloso, 2017; Fonseca et al., 2018) paired with classical machine learning approaches. Additional studies focused on specific competences or aspects of the essays, such as thematic coherence (Passero et al., 2019; Pacheco et al., 2023), punctuation (de Lima

et al., 2023), formal register (Filho et al., 2023), and cohesion (Oliveira et al., 2023).

3 Experimental Setup

3.1 Dataset

The dataset used in the competition consists of 1,235 essays written by 5th to 9th-year students of public schools in Brazil. Each essay is based on a motivating text that accompanies it in the dataset. The essays were annotated by two human evaluators in terms of four competences: formal register, thematic coherence, narrative rhetorical structure, and cohesion. Each competence is scored in a scale of 1 to 5, with higher values indicating better text quality and language proficiency.

For the purpose of the competition, the dataset was split into a training set with 740 samples, a public test set with 125 samples, and a blind test set with 370 samples. The experiments in this study focus on the public test set. The distribution of scores is similar across the training and test sets. However, it is highly unbalanced and, with the exception of the thematic coherence competence, biased towards a single value: 3 for formal register and cohesion and 4 for narrative rhetorical structure.

3.2 Foundation Models

In this study, we explore the use of several foundation models for Portuguese. More specifically, we use the large version of BERTimbau (Souza et al., 2020), which is the most used of such models, and multiple versions of the Albertina PT-* model (Rodrigues et al., 2023), which achieved the state-of-the-art performance on multiple Natural Language Processing (NLP) tasks in Portuguese. We use the two large versions of the Albertina PT-BR model, one trained on brWaC (Wagner Filho et al., 2018) and the other on the OSCAR (Suárez et al., 2019) corpus. Additionally, we use the base version of the Albertina PT-BR model to assess the impact of using a smaller foundation model and the large version of the Albertina PT-PT model to assess the impact of using a foundation model dedicated to a different variety of the language.

3.3 Training & Evaluation

We address the scoring of the essays according to each competence independently as a 5-class classification problem. For each competence, each foundation model is fine-tuned on the training data for 20 epochs. The best epoch is then selected according to the sum of the two evaluation metrics used in the context of the competition: weighted F_1 score and Cohen's κ.

Considering the ordinal nature of the scores, the problem could also be approached as a regression task. However, preliminary experiments revealed a decrease in performance in comparison to the classification approach. Nonetheless, during the prediction phase, in addition to the traditional approach of selecting the class with highest probability, we also explore computing the weighted average of the class probability distribution. This approach, which we refer to as softmax regression, led to more robust predictions in a task of similar nature (Ribeiro et al., 2024).

To account for the non-deterministic aspects of neural approaches and enhance robustness, we performed six independent experimental runs. The evaluation metrics are reported as the average across these runs. All non-error metrics are reported in percentage form.

4 Results

Table 1 shows the average results of our experiments. Comparing the results for the different competences, we can see that the scoring performance is significantly worse for the narrative rhetorical structure than the remaining competences. This was expected, as it presents a more complex problem. Furthermore, the fact that the best results for this competence were achieved using the smaller foundation models suggests that the larger models are overfitting and additional training data is required to capture that complexity.

Looking into the results for the other competences, we can see that using the large Albertina PT-BR model trained on brWaC consistently led to better performance than both BERTimbau, which was trained on the same corpus, and the version of the Albertina PT-BR model that was trained on the OSCAR corpus. Furthermore, we observed significant drops in performance when using the base version of the Albertina PT-BR model, which has one ninth of the parameters of the large versions. For thematic coherence and cohesion, we also observed a drop in performance when using the Albertina PT-PT model. However, it outperformed all the other models for scoring in terms of formal register, in spite of being dedicated to a different Portuguese variety. This is probably due to the fact that it was trained on a large amount

Foundation Model		Formal Register		Thematic Coherence		Rhetorical Structure		Cohesion	
		F_1	κ	F_1	κ	F_1	κ	F_1	κ
BERTimbau Large	CL	70.32	.4434	69.70	.5886	56.83	**.2587**	68.69	.3909
	SR	69.83	.4375	69.39	.5842	56.22	.2442	68.69	.3909
Albertina PT-BR	CL	69.88	.4508	68.66	.5834	53.53	.1777	68.72	.4080
	SR	69.53	.4475	69.70	.5982	54.37	.1920	68.80	.4098
Albertina PT-BR brWaC	CL	72.39	.5115	69.78	.5956	55.37	.2328	**69.88**	**.4306**
	SR	72.24	.5075	**70.29**	**.6079**	55.30	.2265	68.97	.4096
Albertina PT-BR Base	CL	67.79	.4210	66.39	.5464	56.86	.2283	67.96	.3814
	SR	65.85	.3971	66.89	.5534	**56.93**	.2361	67.69	.3776
Albertina PT-PT	CL	73.64	.5222	68.19	.5763	56.20	.2339	67.66	.3738
	SR	**74.07**	**.5308**	67.67	.5720	56.37	.2353	68.15	.3857

Table 1: Average results across the multiple runs. CL stands for classification and SR for softmax regression.

of parliament data, which is typically more formal and better written than generic web-crawled data.

Regarding the prediction approach, the results reveal no clear advantage in using softmax regression, as its impact varies across models. Still, it led to the highest average performance in terms of F_1 for narrative rhetorical structure and both metrics for formal register and thematic coherence.

Finally, it is important to refer that we relied on the models with best performance across all runs to enter the competition. In comparison to the average performance, these represent an improvement between 1 and 4 percentage points in terms of F_1 and between .03 and .09 in terms of agreement.

5 Conclusion

Overall, the results of our experiments show that the agreement between the fine-tuned models for AES and human graders varies between fair and substantial. Thus, although these models can provide cues for essay scoring, significant research is still required towards their improvement, especially for the more complex competences. In this context, as future work, we intend to explore the use of hybrid models that combine the strengths of foundation models with those of manually engineered features specific to each of the competences.

Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

References

Evelin Amorim and Adriano Veloso. 2017. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL): Student Research Workshop*, pages 94–102.

Bruno Smarsaro Bazelato and ECF Amorim. 2013. A Bayesian Classifier to Automatic Correction of Portuguese Essays. In *Conferência Internacional sobre Informática na Educação (TISE)*, pages 779–782.

Hongbo Chen and Ben He. 2013. Automated Essay Scoring by Maximizing Human-Machine Agreement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1741–1752.

Tiago de Lima, Luiz Rodrigues, Valmir Macario, Elyda Freitas, and Rafael Mello. 2023. Automatic Punctuation Verification of School Students' Essay in Portuguese. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 58–70.

Moésio Silva Filho, André Nascimento, Péricles Miranda, Luiz Rodrigues, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Rafael Mello. 2023. Automated Formal Register Scoring of Student Narrative Essays Written in Portuguese. In *Anais do Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)*, pages 1–11.

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically Grading Brazilian Student Essays. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 170–179.

Liam Hannah, Eunice Eunhee Jang, Maitree Shah, and Vinayak Gupta. 2023. Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits. *Language Assessment Quarterly*, 20(4–5):399–420.

Richard H. Haswell. 2006. A Bibliography of Machine Scoring of Student Writing, 1962–2005. In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 17, pages 234–243. Utah State University Press.

Dirk Ifenthaler. 2022. Automated Essay Scoring Systems. In Olaf Zawacki-Richter and Insung Jung, editors, *Handbook of Open, Distance and Digital Education*, pages 1–15. Springer Nature Singapore.

Edmund Jones. 2006. ACCUPLACER's Essay-Scoring Technology: When Reliability Does Not Equal Validity. In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 6, pages 93–113. Utah State University Press.

Vivekanandan S. Kumar and David Boulanger. 2021. Automated Essay Scoring and the Deep Learning Black Box: How are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, 31:538–584.

Sabrina Ludwig, Christian Mayer, Christopher Hansen, Kerstin Eilers, and Steffen Brandt. 2021. Automated Essay Scoring using Transformer Models. *Psych*, 3(4):897–915.

Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2022. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. *Journal of Information and Data Management*, 13(1).

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can Neural Networks Automatically Score Essay Traits? In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 85–91.

Felipe Akio Matsuoka. 2023. Automatic Essay Scoring in a Brazilian Scenario. *Computing Research Repository*, arXiv:2401.00095.

Ken S. McAllister and Edward M. White. 2006. Interested Complicities: The Dialectic of Computer-Assisted Writing Assessment. In Patricia Freitag Ericsson and Richard H. Haswell, editors, *Machine Scoring of Student Essays: Truth and Consequences*, chapter 1, pages 8–27. Utah State University Press.

Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023. Towards Explainable Prediction of Essay Cohesion in Portuguese and English. In *Proceedings of the International Learning Analytics and Knowledge Conference (LAK)*, page 509–519.

Rafael Pacheco, Luiz Rodrigues, Lucas Lins, Péricles Miranda, Valmir Macário, Seiji Isotani, Thiago Cordeiro, Ig Bittencourt, Diego Dermeval, Dragan Gašević, and Rafael Mello. 2023. Automated Thematic Coherence Scoring of Student Essays Written in Portuguese. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1086–1097.

Guilherme Passero, Rafael Ferreira, and Rudimar Luís Scaranto Dazzi. 2019. Off-Topic Essay Detection: A Comparative Study on the Portuguese Language. *Revista Brasileira de Informática na Educação*, 27(3):177–190.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An Automated Essay Scoring Systems: a Systematic Literature Review. *Artificial Intelligence Review*, 55(3):2495–2527.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. *Computing Research Repository*, arXiv:2305.06721.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Workshop on the Challenges in the Management of Large Corpora (CMLC)*, pages 9–16.

Masaki Uto. 2021. A Review of Deep-Neural automated Essay Scoring Models. *Behaviormetrika*, 48(2):459–484.

S. Vijaya Shetty, K. R. Guruvyas, Pranav P. Patil, and Jeevan J. Acharya. 2022. Essay Scoring Systems using AI and Feature Extraction: A Review. In *Proceedings of the International Conference on Communication, Computing and Electronics Systems (ICCCES)*, pages 45–57.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: a New Open Resource for Brazilian Portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4339–4344.

Wilson Zhu and Yu Sun. 2020. Automated Essay Scoring System using Multi-Model Machine Learning. In *Proceedings of the International Conference on Machine Learning Techniques and NLP (MLNLP)*, pages 109–117.