Automatic Punctuation Verification of School Students' Essay in Portuguese

Tiago Barbosa de Lima[1], **Luiz Rodrigues**[3], **Valmir Macario** [1], **Elyda Freitas** [2,4], **Rafael Ferreira Mello** [1,2]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, s/n, Dois Irmãos - CEP: 52171-900 - Recife – PE – Brazil

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)
Rua Cais do Apolo, 220, Recife, - CEP: 50030-390 Recife – PE – Brazil

[3]NEES - UFAL – Av. Lourival Melo Mota, S/N - Cidade Universitária,
Maceió - AL, 57072-970.

[4]Departamento de Sistemas de Informação
Universidade de Pernambuco (UPE) – Caruaru, PE – Brazil

{tiago.blima,valmir.macario,rafael.mello}@ufrpe.br

luiz.rodrigues@nees.ufal.br, elyda.freitas@upe.br

Abstract. *Textual production is a key activity at different levels of education. The analysis of essays encompasses several criteria, such as lexical and syntactic errors, cohesion, and coherence. Within these criteria, how the students include punctuation (i.e., final mark and comma) could influence the quality of the final production. Thus, the literature has proposed several approaches to verifying punctuation correction in students' essays for English. However, despite the advancements in natural language processing models for other languages, there is a significant gap concerning punctuation verification. Therefore, this paper proposed a new approach based on state-of-the-art language models to develop a punctuation prediction method for Portuguese. The proposed model was applied to evaluate the textual productions of students in Brazilian public schools. Finally, the results of this study and its practical implications for educational settings are further discussed.*

1. Introduction

Punctuation is a relevant aspect of learning a new language [Suliman et al. 2019]. The incorrect use of punctuation might lead to a diverse range of miss interpretations [Suliman et al. 2019]. Despite its importance, the literature shows that correct punctuation is a significant problem for pupils and second language learners [Awad 2012]. Therefore, several studies have proposed algorithms and models to analyze the main errors made by pupils and second language learners, aiming to build automatic assistance software to help them [Kurup et al. 2016a]. In this context, Natural Language Processing (NLP) has significantly developed punctuation verification models [Sahami et al. 2011].

Moreover, many grammatical checkers that provide feedback about punctuation errors have been released recently (e.g., Grammarly). However, the problem of automatic punctuation checking is not widely addressed in non-English languages, such as

Portuguese. Despite there being some tools, such as cogroo[1] and language tool[2], none provide enough resources for punctuation verification. Cogroo and Language-tool are able to check minor errors such as sentences punctuated at the beginning or repetitive use of periods or commas in sequence. Therefore, other punctuation misuses as commas separating the subject and verb or the lack of sentences to separate the appositive still lack automatic correction tools. Therefore, exploring punctuation correction for Brazilian Portuguese text is still an open problem in the literature.

The primary approach to analyzing students' essays to detect punctuation errors is to create models that predict the correct punctuation and then compare the outcomes with the students' texts [Suliman et al. 2019, Nagy et al. 2021]. More specifically, the task consists of predicting after which words the punctuation is necessary [Vāravs and Salimbajevs 2018]. Previous papers proposed approaches for punctuation prediction in English [Nagy et al. 2021] and Portuguese [Lima et al. 2022], where the authors evaluated different algorithms (e.g., LSTM, BERT, and Conditional Random Fields (CRF)). Particularly, BERT reached the best results in both cases. Although the promising results of the related paper, to the best of our knowledge, no previous work has applied these models specifically for educational settings.

Furthermore, recently the T5 model [Raffel et al. 2020] has reached more robust results in several NLP tasks, compared to BERT, which achieved state-of-the-art in related work. However, to our best knowledge, T5 has not been used for punctuation prediction yet. Therefore, this paper assesses the performance of BERT and T5 models for punctuation prediction in Brazilian Portuguese within educational texts for elementary school students[Gazzola et al. 2019]. The original dataset was conceived to address automatic readability classification but we adapted the text to punctuation restoration text. We also evaluated the models in a new dataset created containing essays from students in Brazilian public schools and the results reveal that BERT reached better results and generalization for this task. There are promising results from both models, we trained and evaluated the models as punctuation restoration tasks and then we used the model to predict the correct punctuation of the student's essays. The models achieve competitive performance in well-structured sentences, despite a poor outcome in incorrectly written sentences. Finally, the practical implications for education are further discussed such as the main causes for poor performance in non well-structed sentences and possible improvements.

2. Background

This section presents background information on language models and the bottleneck of punctuation analysis in NLP.

2.1. Language Models

Language Models (LM) learn the semantic structure of a specific language from unlabeled text corpora, allowing the automatic creation of relationships between words and sentences. These models might boost several NLP tasks, such as Named Entity Recognition, Text Classification, and Question-Answering (QA) systems [Devlin et al. 2019]. LM gained significant prominence with the development of the BERT model and, later, with T5 [Raffel et al. 2020].

[1]http://www.cogroo.org/

[2]https://languagetool.org/pt-BR

BERT is a language model capable of performing different tasks when fine-tuned. It was first released by [Devlin et al. 2019] in two versions: base and large. Similarly to the English version, [Souza et al. 2020] trained a base from the multi-language checkpoint version of BERT (mBERT) with 110M parameters and large BERT version from the original English version of BERT large with 340M parameters respectively [Devlin et al. 2019, Souza et al. 2020, Devlin 2018].

The literature shows that the pre-trained BERT is suitable for different tasks because they learn deep textual representation [Devlin et al. 2019]. Thus, the concept enables the creation of a diverse range of applications through fine-tuning and minimal architectural changes.

T5 is a text-to-text LM with multi-task capability. It aims to predict a sequence of text [Raffel et al. 2020], differently from BERT, which predicts a single word in a given context. This characteristic allows T5 to perform multiple tasks through text generation, such as text summarization, QA, and translation [Raffel et al. 2020]. Additionally, since the T5 model supports multiple tasks in the same model, one must add a specific tag for each task.

Overall, those robust LMs can carry textual representation to different levels, allowing their application in a diverse range of textual contexts, such as education, health, and justice [Kundu 2021, Perrotta and Selwyn 2020].

2.2. Punctuation analysis

Investigating the use of punctuation is essential to build up strategies for improving students' written communication. There is a significant effort to evaluate grammatical correction systems and punctuation verification systems in the educational context [Kinoshita et al. 2006, Adriaens 1994]. It pushed the development of tutoring systems, data analyses, and other methods to evaluate students' performance. For instance, the Cogroo project can recognize simple punctuation errors [Kinoshita et al. 2006]. On the other hand, there are meaningful advancements on this topic in English with the evaluation of the second language learners and college students [Awad 2012].

In this context, [He 2009] evaluated the performance of a tutoring system for automatic punctuation. The purpose was to provide flag feedback anytime a student does not include mandatory punctuation and suggest improvements on this concern by giving step-by-step instructions to fix multiple errors. The study assessed the performance of the proposed method with ten students, who showed significant improvements in the post-test after using the software considering eight English punctuation rules [He 2009]. Furthermore, the system provides insights into how automatic tutoring software can help students across punctuation challenges [He 2009].

Another work [Nagata and Nakatani 2010] goes even further in analyzing the learning effect of automatic educational software in English. The paper evaluates the impact of precision-oriented and recall-oriented software on learning by comparing the results with a real human tutor. The study analyzed 22 different valid essays of 10 sentences or more made by Japanese college students [Nagata and Nakatani 2010]. The first group wrote without any intervention, the second with human tutoring (4 students), and the third and fourth with precision-orient (6 students) and recall-oriented tutoring system (7 students). The researchers evaluated a grammar corrector system that is closer to a hu-

man tutor when based on precision feedback, where critical errors are detected, but other errors the students must find alone.

In addition to the presented papers, Grammarly is an online tool - also available for mobile, Windows, and MAC - for text corrections powered by an AI engine[3]. The main objective of the Grammarly app is to provide online feedback while the user is typing the tasks, not only for grammatical mistakes but also for punctuation and other tools such as plagiarism. Several studies propose to measure not only the real improvement of punctuation promoted by Grammarly but also the overall student view of the platform through survey questions [Im 2021, ONeill and Russell 2019, Cavaleri and Dianati 2016].

Punctuation plays a crucial role in enhancing text comprehension, necessitating the awareness of NLP models towards punctuation in text prediction tasks [Tilk and Alumäe 2016]. However, certain tasks like Automatic Speech Recognition (ASR) and some models do not predict punctuation correctly in the text [Tilk and Alumäe 2016]. Therefore, the punctuation restoration task aims to utilize machine learning techniques, such as sequence labeling, to automatically predict the missing punctuation [Klejch et al. 2016, Makhija et al. 2019, Nagy et al. 2021].

In general, deep learning models provided the most significant results in the last years when combining pre-training embeddings or using pre-trained models such as BERT for punctuation restoration. For example, the strategy proposed by [Nagy et al. 2021] consists of treating the punctuation restoration problem as a sequence labeling task in which each token receives one of the labels according to the Inside–outside–beginning (BIOS) tagging annotation [Ramshaw and Marcus 1995] where O (no-punctuation) and labels I-COMMA (,), I-PERIOD (.), or I-QUESTION (?), which precedes words with the punctuation. The best performing algorithm of this work obtained an 80.6 F1 score for all labels with the BERT-base developed by [Courtland et al. 2020].

Other works [Nagy et al. 2021, Lima et al. 2022, Tilk and Alumäe 2016, Makhija et al. 2019] used the IWSLT 2012-03 dataset to address punctuation restoration tasks both in English and Portuguese. The IWSLT 2012-03 proposed by [Federico et al. 2012] consists of tedtalks transcriptions in different languages, including Portuguese and English, originally proposed by [Federico et al. 2012] to address Spoken Language Translation (SLT), Speech Recognition and Machine Translation (MT). In turn, the work [Hentschel et al. 2021] adopted another strategy to not only make the punctuation restoration faster but also multitask. They used the ELECTRA model to inject errors in the transcription of the ASR model to make the model more robust. The authors obtained a significant improvement of 11% using a model smaller than BERT. Therefore, punctuation restoration is a widely used strategy to recover punctuation from ASR output, showing significant results in the literature not only in English but also in Portuguese.

Those works provide meaningful insights into the main problems and possible solutions in future works. Since the use of punctuation is a critical evaluation factor for pupils and second language learns, different works evaluate students' punctuation automatically or manually in English [Kurup et al. 2016b, ONeill and Russell 2019,

[3]https://app.grammarly.com/

Im 2021]. However, as far as went our research none of them evaluate the use of punctuation by students in Brazilian Portuguese automatically or manually. Besides, there are only limited tools to address punctuation verification in Brazilian Portuguese text [Kinoshita et al. 2006]. Moreover, the state-of-the-art LM for punctuation prediction tasks (BERT and T5) has not been applied to educational settings. As such, this study proposes the following research questions:

RESEARCH QUESTION 1 (RQ1):
To what extent can BERT and T5 predict the correct punctuation for Portuguese texts?

RESEARCH QUESTION 2 (RQ2):
To what extent can BERT and T5 accurately estimate punctuation errors in students' textual productions?

3. Method

This section presents the datasets, as well as procedures for model selection, assessment, and development adopted in this study.

3.1. Data Description

This study adopted two datasets to train the LMs and evaluate students' punctuation performances when writing essays. The first dataset, named NILC dataset, encompasses a series of school books from different educational levels [Gazzola et al. 2019]. The primary objective of the corpus was to evaluate text complexity. The original NILC dataset includes textbooks focused on elementary, middle, high school, and under-graduated levels. Overall, the dataset consists of 1695 texts and 13016 total sentences. Table 1 describes the number of instances used for training, validation, and testing procedure. The dataset was split using a stratified strategy to maintain the same proportion of both educational levels at training and test.

It is important to highlight that we considered all exclamation marks, semi-colons, and question marks to be periods, similar to both previous works [Nagy et al. 2021, Lima et al. 2022]. Moreover, to the best of our knowledge, we were the first to use this dataset to address punctuation restoration.

Table 1. The final number of texts, sentences, and labels after pre-processing of NILC and MEC datasets.

split	Number of Texts	Number of Sentences	Sentences Elementary I	Sentences Elementary II	I-PERIOD	I -COMMA
train	613	9371	4898	4473	11961	9424
test	597	2604	1361	1243	2621	3335
validation	485	1041	544	497	1424	1044
Total	1695	13016	6803	6213	16006	13803
MEC	256	2004	-	-	2004	1082

The second dataset presented in this paper, called MEC dataset, comprises 265 essays (2004 sentences) by students in middle-school public schools in Brazil. Two expert coders annotated the dataset using three categories: insertion (the student included the

punctuation in the wrong place), missing (the student did not include the required punctuation), and exchange (the student included the wrong punctuation). The coders reached an average agreement of 0.569, according to Cohen's Kappa, which represents a moderate agreement [Landis and Koch 1977]. The dataset encompasses 2004 and 1082 instances of period and comma errors, respectively. As this is a small dataset, it was used only for testing purposes, not for training or validation.

3.2. Model Selection

As detailed in section 2.1, the punctuation restoration is a sequence labeling task. Thus, the language models can infer the results without an additional classification algorithm. In this context, we assessed the performance of BERT and T5 for the problem of punctuation prediction and verification. For the BERT model, we used the Portuguese version released by [Souza et al. 2020] in two different architectures: base (with 110M parameters) and large (with 330M of parameters). It is important to mention that BERT is an encoder-only model that predicts words[Devlin et al. 2019]. Thus, we use it to predict the punctuation directly.

On the other side, the T5 model comprises both encoding and decoding strategies. It means that the T5 architecture allows the use of the same model for different tasks by changing the input tag of input texts [Raffel et al. 2020]. Precisely for this study, we predict the entire sentence, with the corresponding punctuation, aiming to evaluate its correctness. The Portuguese T5 was first released by [Carmo et al. 2020] with four pre-trained models. As the authors recommended, we used the most recent models (i.e., ptt5-base-portuguese-vocab and ptt5-large-portuguese-vocab) with 220M and 760M of parameters, respectively [Carmo et al. 2020].

3.3. Model evaluation

To address RQ1, we assessed the selected models with the NILC dataset using the train, validation, and test split described in table 1. We adopted the evaluation process recommended in the literature [Akbik et al. 2018] to compare the results of the sequential-based models (BERT-based and T5-based).

To evaluate BERT, which does single-word prediction, we applied the traditional NLP evaluation measures used by previous works [Nagy et al. 2021, Lima et al. 2022]: precision, recall, and f-score. In short, precision assesses how accurate the model is in predicting a specific category, while recall measures the number of correctly retrieved instances in the dataset. F1-score is the harmonic mean of both measures, which provides a general performance indicator.

For the T5, which does full sentence prediction, the adequate measure to evaluate is the Bilingual Evaluation Understand (BLEU score) [Papineni et al. 2002]. BLEU captures and evaluates the overlap between the predicted and the reference sentences [Garg and Agarwal 2018]. It has been widely used in the Machine Translation domain for years and was adapted to other tasks, such as QA and Text simplification. After the validation step of the T5 model, we also used precision, recall, and f-score to analyze the results in the test set.

To address RQ2, we selected the best model identified in RQ1 to assess their capability to detect errors automatically in the student-written texts, MEC dataset. In this

case, we decided not to fine-tune the models using the MEC dataset due to the limited number of instances available. Therefore, we measure the performance of the models with precision, recall, and f-score.

3.4. Experimental Setup

We used a google cloud T4 Tesla GPU of 16GB architecture to execute the experimentation. For each model, we evaluated five epochs using the hyper-parameter specified in Table 2, as suggested by [Akbik et al. 2018].

Table 2. Model hyper-parameters for BERT and T5 models.

Parameter	BERT	T5
Learning rate	5.00e-5	5.00e-5
Train batch size	8	2
Eval batch size	8	2
Seed	42	42
Optimizer	Adam with betas=(0.9,0.999) epsilon=1e-08	Adam with betas=(0.9,0.999) epsilon=1e-08
LR scheduler type	linear	linear
Number of epochs	5	5

4. Results

This section presents our results for RQ1 and RQ2.

4.1. RQ1: to what extent can BERT and T5 predict the correct punctuation for Portuguese texts?

The first research question aimed to compare the results of BERT and T5 algorithms with the NILC dataset. Initially, we focused on the analysis of the training and validation process. Figures 1 and 2 present the results of the execution from epochs 1 to 5 in the validation dataset of BERT and T5, respectively. Overall, the best results were reached with four epochs for the base models and five for the large ones. Thus, these were the models selected for the rest of the experimentation.

As can be observed in Figure 1, the detection of the score, comparing the variations of the BERT model, portrays the convergence of the predictive capacity of the model over time. The BERT Base model shows more smoothness in detecting the score and stabilizing itself in constant accuracy with the course of training and validation. The BERT Large model, in contrast, by better capturing phrase-level information in the lower and hierarchical information in the intermediate layers of the language [Jawahar et al. 2019], reaches the highest levels of score prediction.

Unlike the previous scenario, Figure 2 shows the evolution in terms of the BLEU measure of the T5 model. This measure was evaluated to observe the agreement of the model's predicted output with the expected one. It is essential to highlight that the T5 Base model presented the best BLEU measurement in epoch 3. Another essential characteristic is that both T5 Large and Base models presented similar training/validation curves.

4.2. RQ2: to what extent can BERT and T5 accurately estimate punctuation errors in students' textual productions?

Tables 3 and 5 present the comparative results to answer RQ2. That is, under the comparative aspect between the models, to observe Precision, Recall, and F1-Score measures in

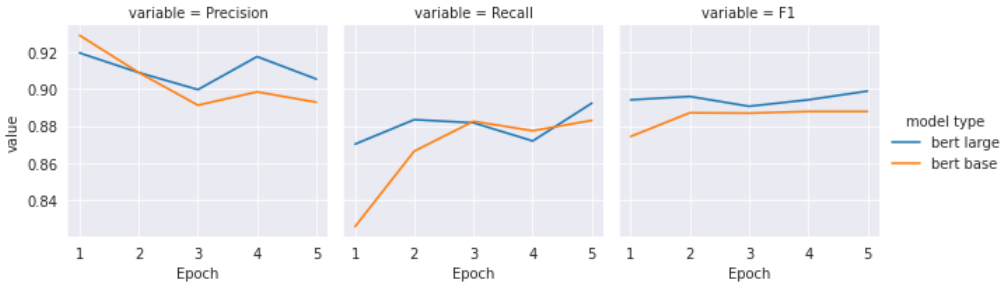

Figure 1. BERT training Evolution with the validation set.

Figure 2. T5 model training performance on the validation set.

both models BERT and T5, considering their Base and Large variants. Hence, the tables enable observing the predictive capacity of the models concerning the evaluation of the punctuation of the texts in the Portuguese Language.

Considering a more controlled dataset, with texts produced and extracted from educational books, Table 3 presents the results obtained by the models and their respective variations. The predictive capacity, in general, was between 0.74 and 0.84 in all measures. Evaluating only *comma* and considering the entire period, the accuracy value rises to 0.98 and 0.99. In terms of the mean, the values obtained were between 0.85 and 0.91. Thus, the best accuracy for evaluating the score, in terms of average, is in the T5 model and its variants BASE and Large, with measures of 0.90 and 0.91, respectively.

Table 3. Table shows the result for all models and measures evaluated at the NILC test dataset with measures Precision (P), Recall (R) and F1-score (F).

	BERT BASE			BERT LARGE		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.802	0.772	0.787	0.81	0.784	0.797
PERIOD	**0.997**	0.993	**0.995**	0.996	0.993	0.994
AVG	0.891	0.873	0.882	0.895	0.88	0.887
	T5 BASE			**T5 LARGE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.831	0.747	0.787	0.842	0.762	0.8
PERIOD	0.995	**0.989**	0.992	**0.998**	**0.994**	**0.996**
AVG	0.906	0.858	0.88	0.914	0.868	0.89

Table 4. Comparison with previous works related to punctuation restoration with measures Precision (P), Recall (R) and F1-score (F).

Paper	Model	Language	P	R	F
[Makhija et al. 2019]	BERT-Punct LARGE	English	79.5	83.7	81.4
[Courtland et al. 2020]	Roberta-base	English	84	**83.9**	**83.9**
[Nagy et al. 2021]	BERT base uncased	English	75.8	85.1	79.8
[Lima et al. 2022]	BERT base cased	Portuguese	83.3	78.9	81

The table shows the previous results of punctuation restoration works on the widely used IWLST2012 public dataset. Differently from our work, the previous paper considered commas, periods and question marks instead of treating question marks as periods.

Table 5 presents the comparative results with the models considering the MEC Dataset. Unlike the previous scenario, which considered the NILC Dataset, the results obtained in this comparison showed a better performance in evaluating the punctuation by the BERT model.

Table 5. Table shows the result for all models and measures evaluated at the MEC dataset with measures Precision (P), Recall (R) and F1-score (F).

	BERT BASE			BERT LARGE		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.12	0.368	0.181	0.123	0.381	0.186
PERIOD	**0.984**	**0.999**	0.991	**0.97**	0.996	**0.983**
AVG	0.707	0.797	**0.732**	0.698	0.799	0.727
	T5 BASE			**T5 LARGE**		
	P	**R**	**F**	**P**	**R**	**F**
COMMA	0.049	0.126	0.07	0.047	0.139	0.07
PERIOD	0.8	0.009	0.018	0.697	0.011	0.021
AVG	0.603	0.04	0.032	0.527	0.044	0.034

Finally, Table 6 presents descriptive statistics of the proportion of errors that were returned. Those consider Different Numbers of Labels (Test case 1), partial evaluation, which corresponds to an Equal Number of Labels but Wrong Placement (Test case 2), and, Full match (Test case 3). That is an approximate assessment of where the probable punctuation error might be. At this stage, some linguistic mechanisms, such as ambiguities, were concentrated, which could result in two possible ways to evaluate the score in the dataset.

Table 6. Number of examples in each case evaluated.

Test Case	Number of Punctuation	Proportion
1	237	54.11%
2	15	3.42%
3	186	42.47%
Total	438	100%

5. Discussion

Punctuation plays a vital role in enhancing the clarity and readability of communication. By providing precise markers, it facilitates effective communication. The results obtained

from our evaluation indicate strong promise when utilizing more recent Natural Language Processing algorithms. Our best result, achieved using the T5-Large algorithm, achieves an impressive average F1-score of 0.89, surpassing the performance of previous work by [Courtland et al. 2020]. Several factors contribute to this positive outcome. Firstly, in our analysis, we considered labels for both periods and commas since the students' datasets treated question marks as periods without analyzing them separately.

As a result, the total number of periods in the training set increases. However, we intentionally refrain from using an excessively large dataset. Doing so may cause the algorithm to overly generalize within a specific context, which differs from the TEDTALK IWLST2012 dataset described in the papers referenced in Table 4. Furthermore, the dataset consists of texts specifically tailored for children, which contributes to achieving a higher level of accuracy. However, it is worth noting that these texts are comparatively simpler compared to more complex and mature content intended for a different audience.

We also address RQ2, which explores the extent to which BERT and T5 models can accurately detect punctuation errors in students' written work. While both models demonstrate above-average performance, particularly BERT, they encounter difficulties likely stemming from the dataset size. The limited number of samples may not provide enough information for the models to effectively evaluate aspects such as pauses, rhythm, and intonation within the text. These aspects are crucial for various text genres like narratives and dissertations. Furthermore, students' essays often contain significant grammatical errors that can introduce punctuation inconsistencies. Language models like BERT, which capture intricate linguistic features at the phrasal level, may face challenges in correctly labeling punctuation, especially when confronted with grammatical mistakes in a sentence [Jawahar et al. 2019]. This can significantly impact the accurate labeling of punctuation, particularly for commas.

Finally, the discussed models demonstrate the ability to verify writing style and provide corrective feedback, showcasing minimum threshold scores that should be present in students' texts. However, a notable limitation of these models is their assessment of commas in student sentences, despite performing well on well-structured sentences. Given the multitude of grammatical errors and text inconsistencies that can lead to erroneous predictions, it would be beneficial to evaluate the maximum number of grammatical errors before assessing punctuation. This approach would help mitigate the problem effectively. Additionally, delving into how the model arrived at a specific label can enhance robustness. Therefore, incorporating explainable AI (XAI) in future research could further improve the results. By building more robust models, we can assist middle-school teachers in essay assessments while boosting students' confidence and enhancing their writing skills [Wilson and Roscoe 2020, He 2009]. Moreover, as one of the pioneering studies addressing automatic punctuation in students' essays, this work opens the door for further research in this area by identifying key limitations and suggesting new directions for investigation. The results highlight that punctuation serves as a valuable tool for evaluating textual continuity, representing intonation and conveying emotions in narrative texts. It underscores the importance of punctuation in writing, enabling more precise, accurate, and effective communication.

6. Final Remarks

Punctuation verification has been addressed in different formats over the year. However, the topic is not fully discussed in Brazilian Portuguese. Thus, this paper presents a benchmark evaluation of BERT and T5 language models to address the punctuation restoration task in Brazilian Portuguese text for children. Also, as far as went our research, no paper to previous data has yet addressed the punctuation verification of students' essays before, then we present a novel dataset for punctuation verification of Brazilian students that can help research in the field in the close future.

The results show that models can be applied with success in well-structured sentences, however, improvements are necessary for unstructured texts. Moreover, punctuation verification with ML has promising results, for future works comparison with ruled-based approaches and LLM prompting engineering would be of good importance.

The results present the evaluation from the local perspective of error correction and its overall relationship shows a strong deficiency in predicting punctuation in a not well-structured text. However, some mechanisms, such as some datasets to emphasize important words and phrases and their due grammatical classes, could be used to enrich the datasets further and, consequently, make the models reach higher levels of score evaluation.

References

Adriaens, G. (1994). Simplified English grammar and style correction in an MT framework: The LRE SECC project. In *Proceedings of Translating and the Computer 16*, London, UK. Aslib.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Awad, A. (2012). The most common punctuation errors made by the english and the tefl majors at an-najah national university. . *Vol.*, 26:23.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Cavaleri, M. R. and Dianati, S. (2016). You want me to check your grammar again? the usefulness of an online grammar checker as perceived by students. *Journal of Academic Language and Learning*, 10(1):A223–A236.

Courtland, M., Faulkner, A., and McElvain, G. (2020). Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 272–279.

Devlin, J. (2018). Multilingual bert readme document.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federico, M., Cettolo, M., Bentivogli, L., Michael, P., and Sebastian, S. (2012). Overview of the iwslt 2012 evaluation campaign. In *Proceedings of the international workshop on spoken language translation (IWSLT)*, pages 12–33.

Garg, A. and Agarwal, M. (2018). Machine translation: a literature review. *arXiv preprint arXiv:1901.01122*.

Gazzola, M. G., Leal, S. E., and Aluísio, S. M. (2019). Predição da complexidade textual de recursos educacionais abertos em português. In *Symposium in Information and Human Language Technology - STIL*. SBC.

He, X. (2009). A web-based intelligent tutoring system for english dictation. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, volume 4, pages 583–586.

Hentschel, M., Tsunoo, E., and Okuda, T. (2021). Making Punctuation Restoration Robust and Fast with Multi-Task Learning and Knowledge Distillation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7773–7777. ISSN: 2379-190X.

Im, H.-J. (2021). The use of an online grammar checker in english writing learning. *Journal of Digital Convergence*, 19(1):51–58.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kinoshita, J., Salvador, L. d. N., and de Menezes, C. E. D. (2006). CoGrOO: a Brazilian-Portuguese grammar checker based on the CETENFOLHA corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Klejch, O., Bell, P., and Renals, S. (2016). Punctuated transcription of multi-genre broadcasts using acoustic and lexical approaches. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 433–440. IEEE.

Kundu, S. (2021). AI in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328.

Kurup, L., Joshi, A., and Shekhokar, N. (2016a). Intelligent Tutoring System for learning English punctuation. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6, Pune, India. IEEE.

Kurup, L., Joshi, A., and Shekhokar, N. (2016b). Intelligent tutoring system for learning english punctuation. In *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, pages 1–6. IEEE.

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

Lima, T. B. D., Miranda, P., Mello, R. F., Wenceslau, M., Bittencourt, I. I., Cordeiro, T. D., and José, J. (2022). Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In *Intelligent Systems: 11th Brazilian Conference, BRACIS*

2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part II, pages 616–630. Springer.

Makhija, K., Ho, T.-N., and Chng, E.-S. (2019). Transfer learning for punctuation prediction. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 268–273. IEEE.

Nagata, R. and Nakatani, K. (2010). Evaluating performance of grammatical error detection to maximize learning effect. In *Coling 2010: Posters*, pages 894–900, Beijing, China. Coling 2010 Organizing Committee.

Nagy, A., Bial, B., and Ács, J. (2021). Automatic punctuation restoration with BERT models.

ONeill, R. and Russell, A. (2019). Stop! grammar time: University students' perceptions of the automated feedback program grammarly. *Australasian Journal of Educational Technology*, 35(1).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Perrotta, C. and Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of ai in education. *Learning, Media and Technology*, 45(3):251–269.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Sahami, M., desJardins, M., Dodds, Z., and Neller, T. (2011). Educational advances in artificial intelligence. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, SIGCSE '11, pages 81–82, New York, NY, USA. Association for Computing Machinery.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Suliman, F., Ben-Ahmeida, M., and Mahalla, S. (2019). Importance of Punctuation Marks for Writing and Reading Comprehension Skills. *(Faculty of Arts Journal)* - , (13):29–53.

Tilk, O. and Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech*, pages 3047–3051.

Vāravs, A. and Salimbajevs, A. (2018). Restoring Punctuation and Capitalization Using Transformer Models. In Dutoit, T., Martín-Vide, C., and Pironkov, G., editors, *Statistical Language and Speech Processing*, volume 11171, pages 91–102. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Wilson, J. and Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.