Using Punctuation Restoration as Punctuation Verification

No Author Given

No Institute Given

Abstract. Punctuation verification is a concern problem for students and second language learns. Since punctuation is meaningful to a precise understand of the text, verify the correct punctuation has gained attention of different works. The mainly works developed tutoring systems to verify the punctuation from students found significant evidences that precision-orient systems play a significant role providing meaningful feedbacks to students of English. However, despite many advancements in the last years, there is a significant gap concern to punctuation verification to Brazilian Portuguese. Therefore, in this paper, we investigate different methods using language models to punctuation verification. We use punctuation restoration as an ancillary task by restoring non-punctuated task and comparing the results with the student given punctuation. We present a new pupils dataset for punctuation verification, both language show promising results.

Keywords: Punctuation Verification · Language Model · Artificial Intelligence.

1 Introduction

Punctuation is a significant part of learning process of a new language [27]. The incorrect use of punctuation might leads to a diverse range of miss interpretation[27]. Thus, correct punctuate is a significant problem for pupils, and second language learns[2]. For that reason, different researches moved a step forward to analyses the mainly mistakes by pupils and second language learns and building automatic assistance software to help them. Furthermore, Artificial Intelligence (AI) has played a major role on this area as part of automatic grammatical corrector, punctuation versifiers and others [23]. Therefore, in this paper, we present how pre-trained models might be useful to address the problem of automatic punctuation verification. There are still challenge and improvement, but there are promising results.

The work [2] evaluated the impact of punctuation misuse of 100 major university students in English. The result found that most part of them overuse the comma, and also has significant misuse of capital letters [2]. The research evaluate also found that there not a significant gender contrast, however, academic level played a noteworthy role on the number of mistakes [2]. Further, the work motivated the creation of strategies to learn English punctuation at the

university department and improve the written skills of the students [2]. Thus, investigate the use of punctuation is essential to build up strategies of how to improve written communication of students. Further, that provides significant insights to build useful and effective Machine Learning (ML) educational software's to assist orthography, grammar and punctuation correction.

The necessity of providing feedbacks for a wide number students in a effective way inspired the development of educational software's. One of the concerning areas is the punctuation verification of written text as investigated by the work [10] in which a intelligent tutoring system provided feedbacks for punctuation corrections automatically. The research identified that the one of the major drawbacks was the sentence generation process that turned the learning disengaging for students [10]. Then, the research build a specific sentence generator following the subject-verb-object order besides a punctuation system based on rules [10]. The works using tutoring system shows that it high potential of use in different context with a significant impact on the learning process. Moreover, it demonstrate that the precision-oriented feedback systems approximate better to a human tutor. However, the most part of the test were conducted in a small scale scenario and do not extensive use ML and Artificial Intelligence algorithms to empower them. Therefore, there is a significant gap on the provision of capable software to provide useful feedbacks to the learning activity not only in terms of methods, but also in terms of language since most part of works focus on English.

Despite the advances on pre-trained models and methods, ML and AI has not been extensively used in the works [10, 2, 16]. Meanwhile, the works [4] released Bidirectional Encoder Representation (BERT) model capable of achieving state-of-the-art results in different comprehensive tasks, for instance, Question Answering (QA), sentence similarity, Named Entity Recognition (NER) and others [4]. It has boosted the results of many different applications and in different context by fine-tuning a huge model in a specific domain and task [12, 24]. The exactly concepts applies to T5 model developed by [22], a more robust model able to train and make inference in different tasks with the same model boosting semantic transfer learning between them. Thus, there are significant important opportunities to use robust language models to educational area through fine-tuning in downstream tasks and domain specific datasets. Besides, it is significant to search for ways to adapt different concepts to make punctuation verification and orthography checking possible.

Therefore, this paper presents a initial study about the use of BERT and T5 model to punctuation verification. The study uses two versions of each model (one base and one large) and two different datasets, one for training, test and validation and other to test-case analyses only. There are promising results from both models uses in well-structured sentences despite a poor outcome in a incorrect written sentences. The next step is to build and evaluate more noise robust models to punctuation verification and grammatical analyses.

2 Preliminaries

There are several application of AI system nowadays, it includes medicine, automation, education and others [9, 21, 32]. At education AI application range in different ways specially to essay correction, grammatical analyses and others [15, 28]. The applied methods range from classical ML algorithms such as Naïve Bayes, support vector machines and others to more robust methods such as Bidirectional Long-short Memory (LSTM), Transformers and others [5, 25, 13]. In Portuguese, a diverse range of papers focus on the evaluation of essays and orthography mistakes applying a range of algorithms [8]. In this paper, however, we will use punctuation restoration as a ancillary task to punctuation evaluation in educational texts. That consists of predicting words where a punctuation is necessary what was part of different works [31]. The problem was addressed in Portuguese by the work [14] that evaluated three algorithms, LSTM, BERT and Conditional Random Fields (CRF). The results shows that BERT model outperforms other methods in the final results and shows to be have robust performance in out-of-domain application [14]. Language Models are a specific category of models able to learn textual representation by seeing an amount of corpora of text [4]. That allow the automatic creation of relation between words and sentences what might boost several downstream task such as NER, QA and others [4]. Language models gained significant prominence with the development of BERT model and later with T5 [22]. Those robust language models are capable of carrying textual representation to different levels allowing application in a diverse range of textual context, such as Medical texts, legal texts and others [9, 21]. Therefore, we apply BERT and T5 models to punctuation restoration using a sequence labelling approach similar to [19, 14] and other works. Thus, the model will be able to recognise missing punctuation in educational texts allowing the creation of a system to automatic correct miss punctuated text.

Language models were recently boosted due to the concept of attention based transformers [30]. The attention mechanism first introduced by [30] allowed models to extract features from specific parts of the sentences, a group of tokens or a single token, instead of the whole sentences. Attention mechanism improved translation task primarily but was also used in other areas such as visual AI. Later, the paper [30] lunched the attention only mechanism dispensing the use of high computational cost of LSTM layers. The cost reduction allow the training and evaluation of larger models such as BERT and T5 [4, 22]. The attention only mechanism core found is the use of triple of Query (Q), Values (V) and Keys (K) vectors to define a attention matrices to input tokens [30]. The Q is the vector which will select V tokens using specific Keys what is depicted in the equation [30]. Therefore, the final matrix is a mask two select a defined range of desired inputs [30]. That permitted a massive cost reduction enabling a range of different application and the creation of massive language models [30].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{d_k}\right) \tag{1}$$

2.1 Bidirectional Encoder Representation (BERT)

Bidirectional Encoder Representation (BERT) is a language model capable of performing different tasks when fine-tuned. It was first released by [4] in two versions base and large. The purpose is to train a model able to predict masked such as first proposed by [29] and showed by figure 1. It requires a huge amount of data usually grap from web scrapping sources [4]. That allows the model to understand textual representation what makes it suitable to several applications [4]. Therefore, pre-trained model masked models are suitable to different task because the learn deep textual representation to making and perdition techniques. Thus, the concept enables the creation of a diverse range of application through fine-tuning and minimal architectures changes.

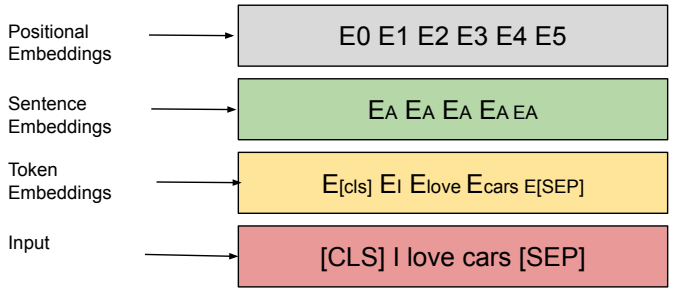

Fig. 1. The image shows the process of embbeding a two sentences using BERT. [SEP] is a special token for sentence separation and [CLS] indicates the beginning of one sentence [4]

2.2 T5

T5 is a text-to-text language model with multi-task capability. T5 model first released by [22] is textual model that differently form BERT where the objective is to predict a single word given a context, the objective is to predict multiple words simultaneously. That permitted the creation of a model able to perform multiple task through text generation such as text summarization, QA, translation and others [22]. T5 uses the same transformer attention based architecture with some slight different from previous BERT, but instead of a encoder-only modelling architecture, T5 is designed for both encoding and decoding. Additionally, since T5 model support multiple tasks in the same model, one needs to add a specific tag specific for each task, for example, translation task always

starts with 'translate English to France:' as exemplified by figure 2. Since T5 was trained with a massive textual corpus, it developed the capability to perform several linguistic tasks through a simple fine-tuning just as BERT [22]. Therefore, we chose to use to the sequence labelling task. Since NER was the closest task we decided to use the same tag label 'Name Entities' when building the dataset.

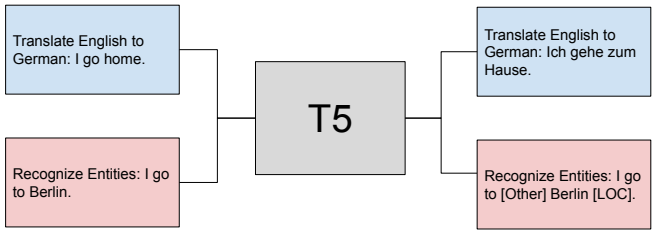

Fig. 2. T5 model multi-task capability with translation and entity recognition.

3 Related Works

There is a significant effort to evaluate grammatical correction system in the educational context [8, 1]. It pushed the development of tutoring systems, data analyses and other methods to evaluate students performance. Cogroo project, for instance, provides a little all no support to punctuation verification considering when punctuation is in front of the sentence and repeated [1]. Also, multilanguage tools such as language tool provides little support to punctuation verification with performance similar to cogroo [8]. By the other hand, there are meaningful advancement on the this topic in English with evaluation of second language learns and college students [2]. Despite the no application of AI and ML, those works provides meaningful insides about the main problems and possible solutions to address in future works. Therefore, there is a significant gap on punctuation verification in Portuguese, however, English related works provides a significant inside about the topic that this paper will address.

One of the most important steps when dealing with educational problem is the analyse the main problems through a field research. It is significant to find most common errors and successful steps to follow in future researches. The work [2] analysed the performance of English student learns at the An-Najah

[1] https://languagetool.org/pt-BR, https://cogroo.sourceforge.net/

National University to analyse students performance according the punctuation errors. The objective to analyse the main students errors, evaluate significant differences according gender, education level and department [2]. The study conduct with 45 man and 55 woman Arab students found that the most common errors are the misuse of comma and period [2]. Some researchers concluded that the reason was because Arab language differentiated significantly from English inducing some of the mistakes [2]. Further, despite gender has no significant influence in punctuation, academic level played a significant role by minimising the errors as higher as it is [2]. Therefore, the study points the main errors and characteristics of punctuation mistakes of English second speakers in the Arab context [2]. It is evident the academic level shows a meaningful difference in the academic year level pattern. It provides significant insides to future works and intelligence systems development in this area.

Evaluate automatic corrections systems is a important step to collect users feedbacks and measure the system effectiveness. This was the purpose of [7], the paper evaluates the performance of a tutoring system for automatic punctuation. The purpose was to flag anytime a student forget a mandatory punctuation and suggest improvements on this concern given step by step instruction to fix multiple errors [7]. The test made with 10 test showed significant improvement in the post-test after the use of the software considering 8 English punctuation rules [7]. Further, the system provides insights that automatic tutoring software can definitely help students across punctuation challenges [7]. Another work [18] goes even further in the analyses of the learning effect of automatic educational softwares in English. The paper evaluates the impact of precision-oriented and recall oriented softwares in the learning process comparing the results with a real human tutor. The study conducted analysed valid 22 in different essay of 10 sentence or more made by Japanese college students [7]. The first group wrote without any intervention 5, the second with human tutoring (4 students) and third and fourth with precision-orient (6 students) and recall-oriented tutoring system (7 students) [7]. The researches evaluate a grammar corrector system is more close to a human tutor when based on the precision feedback where critical errors are detected, but others a the students must find alone.

Finally, investigating the system performance of the model in a wider context is also important. Thus, it is noteworthy to evaluate along with other criterion and in other contexts such as in essay evaluation, maximisation of the learning effect and the errors evaluation and analyses. This provided useful insights to this research and we therefore, we present a analyses of a Brazilian corpus for punctuation evaluation of pupils students. We analyses the main errors and provided feedbacks about the educational dataset. Further, we build a synthetic dataset for punctuation restoration that we will be used as an ancillary task.

4 Materials and Methods

We use punctuation restoration as an ancillary task for punctuation correction by following a NER strategy that is widely used in literature to fix ASR output. The

purpose is to remove all punctuation from the text and then predict all of them and finally compare the punctuation with the one provided by the students. It will allow us to automatically correct missing punctuation or incorrect insertion of punctuation by students.

4.1 - Datasets

The first dataset is a NILC dataset of school books of different educational levels available by [17]. The primary objective of the corpus was to train and evaluate different corpus according to the educational level they belong aiming to curate educational resources. The second one is an annotated school pupil dataset of different topics with a few thousand sentences and more than 200 hundred texts. Since the second dataset has poor quality due to the educational level of the kids and it does not have many examples we only use it for evaluation purposes not considering it in the training pipeline. Thus, each word followed by a punctuation mark (i.e ',' or '.') is labelled as I-COMMA or I-PERIOD as shown in table 4.1.

Você	vai	para	casa	não	é
O	O	O	I-COMMA	O	I-PERIOD

First, we use models evaluating each one in this task, and afterwards we consider Brazilian students' essays of elementary school students according to the model's prediction.

NILC Dataset: The corpus has for splits, I-Elementary (Fundamental I), II-Elementary (Fundamental II), High School (Ensino Médio), and under-graduated (Ensino Superior) composed of textbooks, WikiBooks, text from the National High School Exam (ENEM) from 2015 to 2017 and news text re-writing to kids aged between 8 and 11. To train and evaluate the method, we chose to use only the text from the I and II elementary years due to the higher complexity of other texts compared to the ones used to evaluate punctuation. Therefore, the dataset consists of 1695 texts from sources and 13016 total sentences. The process of dataset construction is depicted in the figure 3 [2].

Students Essay Dataset: The educational dataset is composed of 2004 sentences and 265 different texts written by students at the elementary school level. The consists of written essays that were annotated by two different annotators and includes different grammar, orthography and punctuation evaluations. Here we only consider the punctuation evaluation for instance, the table 4 resumes the number of labels for both I-COMMA and I-PERIOD. Finally, the table 3 shows the statistics of erros in the MEC test dataset and the table 4 shows the statistics.

[2] We used sentence tokenize provided by NLTK toolkit that does not always consider period as sentences. Example: The Mr. John went home. There are two periods but only one sentences.

Table 1. The proportion of educational level dataset.

Education Level	General Proportion
Ensino Fundamental I	16.4
Ensino Fundamental II	14.9
Ensino Medio	30.1
Ensino Superior	38.7

Fig. 3. The dataset construction pipeline. 1. The data prepossessing including removing unwanted characters (e.g ¡¿ {}[]) normalizing all the punctuation replacing '!' '?' and ';' for PERIOD and splitting the text in sentences. 2. we transform the data by splitting each piece of text in tokens. 3. We filter to only include Elementary I and II.

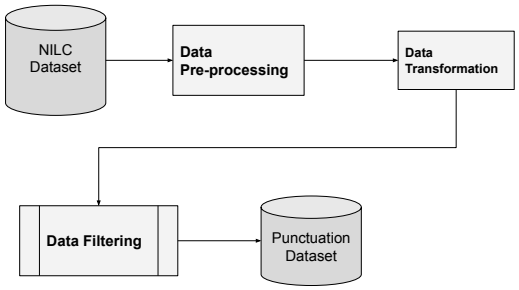

Table 2. The final number of texts, sentences and labels after pre-processing. We split the examples using a stratified strategy to maintain the same proportiong both educational level at training and test.

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

Table 3. The dataset shows the number and types of errors in the MEC test dataset.

both annotators		
	I-PERIOD	I-COMMA
insertion	19	80
missing	1205	107
exchange	49	0

Since the first dataset is used during training, this second one is used for evaluation purposes only. The manuscript was annotated by two annotators and the annotation was merged such that each text has all possible punctuation required. Further, it is important to highlight that since those manuscripts are annotated by students in their first years of school significant grammatical and

Table 4. MEC dataset statistics

Dataset - Labels	Number
O	28683
I-PERIOD	2004
I-COMMA	1082
Total	36432

orthography mistakes might be found which has a significant negative impact on model prediction as we will show later in section 5.

4.2 Methods

We decided to separate the work in 4 dissimilar steps to train and evaluate the models in different scenarios. Thus, our strategy follows the steps:

1. Training punctuation restoration models for commas and periods.
2. Evaluate the models at the original domain text.
3. Verify the model performance at a dataset of elementary school students' essays.
4. Discuss different study case scenarios and further improvements.

Firstly, we chose the punctuation restoration strategy proposed by [19] and also evaluate in Brazilian Portuguese by [14]. It consists of removing all the punctuation from the text and then predicting the punctuation using a machine learning model. Deep learning models provided the most significant results in last years when combining pre-training embeddings or using pre-trained modes such as BERT. The strategy proposed by [19] and followed by consists of treating punctuation restoration problem as a sequence labelling task which each token is receives one of the labels O, I-COMMA or I-PERIOD. I-COMMA labels words (or tokens) that precedes a comma mark I-PERIOD a period and O means Other or no-punctuation. Both [19] and [14] considered three punctuation I-COMMA, I-PERIOD and I-QUESTION, however, since our one of the test datasets consists only the two first ones we decided to normalise all exclamation marks, semi-colon and question marks to periods similarly to [19, 14]. Then, we tested and evaluated the model in 4 different models depicted at subsection 4.3.

Secondly, we tested the model considering the punctuation restoration at the original domain which the model was trained. It provides useful insights about which model might perform better when considering well-structured sentences of texts. Finally, we evaluate the models at the a dataset of elementary school students with poor standardisation of the written language and several grammatical and orthographic mistakes that might impact the models performance. It might demonstrate us the models robustness when dealing with poor written texts, as well as, will demonstrate the transfer learning capability. At end, we show 4 contrasting situations where the model might correctly predict the punctuation, incorrect predict the punctuation or partially provide the correct outcome.

4.3 Models

Firstly, we chose to train and evaluate 4 different models and two different architectures. The first type of model is the BERT model proposed by [4] which is extensively used for many different tasks such as NER, QA, Topic Modelling and others [11]. Since BERT is an encoder-only model it is suitable for text comprehension tasks such as QA and NER which are specialized at extracting meaningful information from the input text. We used the Portuguese version released by [26] in two different architectural: base with 110M and large with 330M of parameters both of them considering case words, the table 5 depicts both architectures. The second type of model is a t5 model that comprises both encoding and decoding strategies. The major advance of T5 architecture is to allow us to use the same model for different tasks by changing the input tag of input texts. It permits the extraction of full cross-knowledge between two dissimilar tasks, but that together provide significantly better results [22]. Further, it might be useful to train and evaluate a single model instead of training different models. The Portuguese t5 was first released by [3] with 4 ptt5-small-t5-vocab, ptt5-base-t5-vocab, ptt5-large-t5-vocab, ptt5-small-portuguese-vocab, ptt5-base-portuguese-vocab (Recommended), ptt5-large-portuguese-vocab. We chose the last two versions ptt5-base-portuguese-vocab (Recommended), ptt5-large-portuguese-vocab with 220M and 760M of paramters respectively.

BERT models The BERT architecture has been widely used for solving NER problems. Different works use propose fined-tuned BERT for NER tasks. In the work [14], the bert model achieves significant results in the task of punctuation restoration in the English version of IWLST tedtalk dataset in Portuguese. Based on this approach we evaluate the model performance.

Table 5. BERT architecture

BERT Architecture			
LARGE		**BASE**	
vocab size	29794	vocab size	29794
hidden size	1024	hidden size	768
num hidden layers	24	num hidden layers	12
num attention heads	16	num attention heads	12
hidden act	gelu	hidden act	gelu
intermediate size	4096	intermediate size	3072
hidden dropout prob	0.1	hidden dropout prob	0.1
attention probs dropout prob	0.1	attention probs dropout prob	0.1
max position embeddings	512	max position embeddings	512
type vocab size	2	type vocab size	2
initializer range	0.02	initializer range	0.02
layer norm eps	1.00E-12	layer norm eps	1.00E-12
position embedding type	absolute	position embedding type	absolute

T5 Architecture T5 architecture is widely used for multi-task development models. It same model can perform different tasks dispensing the training of different architecture. It has also achieved successful results at NER task. The pre-trained model in this paper was released by [4] and evaluate at the HAREM dataset for NER task achieving comparable results with SoTA NER BERT model.

Table 6. T5 architecture

T5 Architecture			
LARGE		**BASE**	
Vocab size	32128	vocab size	32128
Dimension model	1024	d model	768
Num. of positions	512	n_positions	512
Dimension feedfoward	4096	Dimension feedfoward	3072
Num. layers	24	num layers	12
Num. decoder layers	24	num decoder layers	12
Num. heads	16	num heads	12
Relative attention num buckets	32	relative Attention num buckets	32
Relative attention max distance	128	relative Attention max distance	128
Dropout rate	0.1	dropout rate	0.1
Layer norm epsilon	1.00E-06	layer norm epsilon	1.00E-06
feed forward proj	relu	feed forward proj	relu
dense act fn	relu	dense act fn	relu

4.4 Evaluation Metrics

The evaluation metrics used to evaluate the performance during the BERT model training were Recall, Precision, F1-Score. Firstly, Recall is the metrics that measures how well the model can retrieve (predict) labels that should be predicted. Thus, according to our scenario, we will able to evaluate the proportion of I-PERIOD and I-COMMA that the model are true negatives and false negatives. Secondly, Precision measure how well the model predict true positive labels, labels that are correctly predicted given us the proportion of true positives versus false positives. Finally, F1-score is the balanced between both metrics which might be good indicator of how the performance goes in general. Therefore, we used those metrics to evaluate the training performance of BERT model at training and test set and the T5 model at the test set.

Another important metrics is the Bilingual Evaluation Understand (BLEU score) that measure how a model generated strings matches the expected output [20]. Its widely use at Machine Translation (MT) is domain for years and it was adapted to other task such as QA and Text simplification. It captures and evaluates the overlaps between the predicted token and the reference token [6]. As bigger the overlap is as high the score. Thus, we used Bleu score to measure the performance of the prediction during training of the T5 model [20].

5 Results

5.1 Experiment Setup

The experiments showed the models able to converge to optimal results with just a few epochs of fine-tuned training. First, BERT models in both configurations achieve considerable results better than the t5 model. We ran each model for 5 epochs, each one with the specified hyper-parameter at table 7 on a T4 Tesla GPU of 16GB.

Table 7. Model hyper-parameters for BERT and T5 models.

Training Hyper-parameters		
Parameter	BERT	T5
learning rate	5.00E-05	5.00E-05
train batch size	8	2
eval batch size	8	2
seed	42	42
optimizer	Adam with betas=(0.9,0.999) and epsilon=1e-08	Adam with betas=(0.9,0.999) and epsilon=1e-08
lr scheduler type	linear	linear
num epochs	5	5

5.2 Results

The training show that both models types achieve comparable results as shows the graphs5 and 4 with only a few epochs.

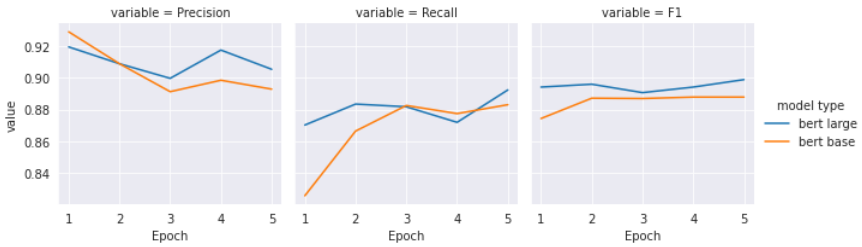

Fig. 4. BERT training Evolution with the validation set.

It is also true for the test set where the results shows the high performance of all models with special attention to T5 models that achieves the highest precision score (T5 LARGE) and receives that highest F1 score (T5 BASE).

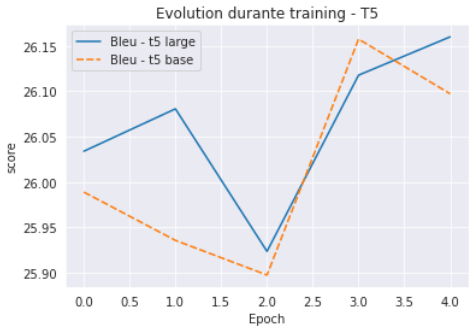

Fig. 5. T5 model training performance on the validation set.

Table 8. Table shows the result for all models and metrics evaluated at the NILC test dataset.

	precision	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
	BERT BASE			**BERT LARGE**		
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
micro avg	0.893	0.873	0.883	0.896	0.88	0.888
macro avg	0.9	0.883	0.891	0.903	0.889	0.896
weighted avg	0.891	0.873	0.882	0.895	0.88	0.887
	T5 BASE			**T5 LARGE**		
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
micro avg	0.91	0.858	0.883	**0.917**	0.868	0.892
macro avg	0.913	0.868	0.889	**0.92**	0.878	0.898
weighted avg	0.906	0.858	0.88	**0.914**	0.868	0.89

5.3 Test Cases Evaluation

We also evaluated the results in terms of prediction quality. We extracted a sample of 438 predictions 20% from the total. We splited the predictions in three different categories: full match, which means that the prediction full matched with the reference labels when accuracy is above 99%, in the second category are prediction that the number of punctuation are correct, but they are placed in the wrong place and finally, the last category of prediction considers prediction which the number of labels is different from the references ones.

6 Conclusion

Punctuation verification has been addressed in different format along the years. However, there is a clear gap in the literature concern to the topic where no paper as long as went our research evaluate the performance of Machine Learning algorithms in punctuation verification. Also, the topic is not fulled discussed in

Table 9. Table shows the result for all models and metrics evaluated at the MEC dataset.

	BERT BASE			**BERT LARGE**		
	precision	recall	f1-score	precision	recall	f1-score
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
micro avg	0.477	0.797	0.597	0.472	**0.799**	0.593
macro avg	0.552	0.684	0.586	0.546	**0.688**	0.584
weighted avg	0.707	0.797	0.732	0.698	0.799	0.727
	T5 BASE			**T5 LARGE**		
	precision	recall	f1-score	precision	recall	f1-score
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
micro avg	0.058	0.04	0.047	0.056	0.044	0.05
macro avg	0.424	0.068	0.044	0.372	0.075	0.046
weighted avg	0.603	0.04	0.032	0.527	0.044	0.034

Table 10. Number of examples in each case evaluated.

Test Case	Number of Punctuation	Proportion
Different Number of Labels	237	54.10958904
Full Match	186	42.46575342
Equal Number of Labels but Wrong Placement	15	3.424657534
Total	438	100

Table 11. Examples of prediction when different number of punctuation are predicted in comparison to the reference.

Different Number of Labels	
Prediction	e depois que a chuva ***baixou,*** encontrei no quintal da minha casa uma pedra muito **brilhante.**
Ground truth	e depois que a chuva baixou encontrei no quintal da minha casa uma pedra muito **brilhante.**
Prediction	dobrei as roupas e fui para a escola ***e,*** quando ***encasa,*** achei uma **chave.**
Ground truth	dobrei as roupas e fui para a escola e quando encasa achei uma **chave.**
Prediction	e eu pedi ajuda para minha mãe e ela arrumou para **mim.**
Ground truth	e ***eu,*** pedi ajuda para minha mãe e ela arrumou para **mim.**

Brazilian Portuguese. Thus, this paper presents a new dataset for punctuation verification considering brazilian students pupils. Further, the results shows that models can be applied with success in well-structed sentences, however, improvements are necessary for unstructured texts. Moreover, punctuation verification

Table 12. Examples of where we have full match with accuracy bigger than 99%.

Full Match	
Prediction	e ela descobriu que as cores são maravilhosas.
Ground truth	e ela descobriu que as cores são maravilhosas.
Prediction	a menina perguntou para todo mundo da família.
Ground truth	a menina perguntou para todo mundo da família.
Prediction	eli dechou a genti brinca nela na barca gigante.
Ground truth	eli dechou a genti brinca nela na barca gigante.
Prediction	como essa arvore veio parar aqui.
Ground truth	como essa arvore veio parar aqui.

Table 13. Prediction with the same number of punctuation, but not full match.

Equal Number of Labels but Wrong Placement	
Prediction	numa ***tarde,*** uma menina viu ao tintas no balcão e pegou as tintas e foi **brincar.**
Ground truth	numa tarde uma menina viu ao tintas no ***balcão,*** e pegou as tintas e foi **brincar.**
Prediction	fiquei com duas toda vez que ***chovia,*** as vezes caia pedras **brilhosas.**
Ground truth	fiquei com **duas,** toda vez que chovia as vezes caia pedras **brilhosas.**
Prediction	não gostava de ***nada,*** ele não gostava de ***brincar,*** ele.
Ground truth	não gostava de **nada, ele,** não gostava de brincar **ele.**

with ML has promising results, for future works comparasion with ruled based approaches would be of good importance.

References

1. Adriaens, G.: Simplified english grammar and style correction in an mt framework: the lre secc project. In: Aslib proceedings. MCB UP Ltd (1995)
2. Awad, A.: The most common punctuation errors made by the english and the tefl majors at an-najah national university. . Vol. **26**, 23 (2012)
3. Carmo, D., Piau, M., Campiotti, I., Nogueira, R., Lotufo, R.: Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., Gasevic, D.: Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. In: LAK22: 12th International Learning Analytics and Knowledge Conference. pp. 404–414. ACM, Online USA (Mar 2022). https://doi.org/10.1145/3506860.3506977, https://dl.acm.org/doi/10.1145/3506860.3506977

6. Garg, A., Agarwal, M.: Machine translation: a literature review. arXiv preprint arXiv:1901.01122 (2018)
7. He, X.: A web-based intelligent tutoring system for english dictation. In: 2009 International Conference on Artificial Intelligence and Computational Intelligence. vol. 4, pp. 583–586 (2009). https://doi.org/10.1109/AICI.2009.37
8. Kinoshita, J., do Nascimento Salvador, L., de Menezes, C.E.D.: Cogroo: a brazilian-portuguese grammar checker based on the cetenfolha corpus. In: LREC. pp. 2190–2193 (2006)
9. Kundu, S.: Ai in medicine must be explainable. Nature medicine **27**(8), 1328–1328 (2021)
10. Kurup, L., Joshi, A., Shekhokar, N.: Intelligent tutoring system for learning english punctuation. In: 2016 International Conference on Computing Communication Control and automation (ICCUBEA). pp. 1–6 (2016). https://doi.org/10.1109/ICCUBEA.2016.7860019
11. Labusch, K., Kulturbesitz, P., Neudecker, C., Zellhöfer, D.: Bert for named entity recognition in contemporary and historical german. In: Proceedings of the 15th conference on natural language processing. pp. 9–11 (2019)
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
13. Liang, G., On, B.W., Jeong, D., Kim, H.C., Choi, G.S.: Automated essay scoring: A siamese bidirectional lstm neural network architecture. Symmetry **10**(12), 682 (2018)
14. Lima, T.D., Miranda, P., Mello, R.F., Wenceslau, M., Bittencourt, I.I., Cordeiro, T., José, J.: Sequence labeling algorithms for punctuation restoration in brazilian portuguese texts. In: BRACIS 2022 () (nov 2022), http://XXXXX/226026.pdf
15. Marinho, J., Anchiêta, R., Moura, R.: Essay-br: a brazilian corpus to automatic essay scoring task. Journal of Information and Data Management **13** (2022)
16. Mayo, M., Mitrovic, A., McKenzie, J.: CAPIT: an intelligent tutoring system for capitalisation and punctuation. In: Proceedings International Workshop on Advanced Learning Technologies. IWALT 2000. Advanced Learning Technology: Design and Development Issues. pp. 151–154. IEEE Comput. Soc, Palmerston North, New Zealand (2000). https://doi.org/10.1109/IWALT.2000.890594, http://ieeexplore.ieee.org/document/890594/
17. Murilo Gazzola, Sidney Evaldo Leal, S.M.A.: Predição da complexidade textual de recursos educacionais abertos em português. In: Proceedings of the Brazilian Symposium in Information and Human Language Technology (2019)
18. Nagata, R., Nakatani, K.: Evaluating performance of grammatical error detection to maximize learning effect. In: Coling 2010: Posters. pp. 894–900. Coling 2010 Organizing Committee, Beijing, China (Aug 2010), https://aclanthology.org/C10-2103
19. Nagy, A., Bial, B., Ács, J.: Automatic punctuation restoration with BERT models (Jan 2021)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
21. Perrotta, C., Selwyn, N.: Deep learning goes to school: Toward a relational understanding of ai in education. Learning, Media and Technology **45**(3), 251–269 (2020)

22. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
23. Sahami, M., desJardins, M., Dodds, Z., Neller, T.: Educational advances in artificial intelligence. In: Proceedings of the 42nd ACM technical symposium on Computer science education. pp. 81–82. SIGCSE '11, Association for Computing Machinery, New York, NY, USA (Mar 2011). https://doi.org/10.1145/1953163.1953189, https://doi.org/10.1145/1953163.1953189
24. Silveira, R., Fernandes, C., Neto, J.A.M., Furtado, V., Pimentel Filho, J.E.: Topic modelling of legal documents via legal-bert. Proceedings http://ceur-ws org ISSN **1613**, 0073 (2021)
25. Sousa, A., Leite, B., Rocha, G., Lopes Cardoso, H.: Cross-Lingual Annotation Projection for Argument Mining in Portuguese. In: Marreiros, G., Melo, F.S., Lau, N., Lopes Cardoso, H., Reis, L.P. (eds.) Progress in Artificial Intelligence, vol. 12981, pp. 752–765. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86230-5$_5$9, $https$: $//link.springer.com/10.1007/978 - 3 - 030 - 86230 - 5_5 9, seriesTitle$: $LectureNotesinComputerScience$
26. Souza, F., Nogueira, R., Lotufo, R.: Bertimbau: pretrained bert models for brazilian portuguese. In: Brazilian conference on intelligent systems. pp. 403–417. Springer (2020)
27. Suliman, F., Ben-Ahmeida, M., Mahalla, S.: Importance of Punctuation Marks for Writing and Reading Comprehension Skills. (Faculty of Arts Journal) - (13), 29–53 (Jun 2019). https://doi.org/10.36602/faj.2019.n13.06, https://misuratau.edu.ly/journal/arts/upload/file/R-404-8.pdf
28. Tashu, T.M., Horváth, T.: Synonym-based essay generation and augmentation for robust automatic essay scoring. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 12–21. Springer (2022)
29. Taylor, W.L.: "cloze procedure": A new tool for measuring readability. Journalism quarterly **30**(4), 415–433 (1953)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Vāravs, A., Salimbajevs, A.: Restoring Punctuation and Capitalization Using Transformer Models. In: Dutoit, T., Martín-Vide, C., Pironkov, G. (eds.) Statistical Language and Speech Processing, vol. 11171, pp. 91–102. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00810-9$_9$, $http$: $//link.springer.com/10.1007/978 - 3 - 030 - 00810 - 9_9, seriesTitle$: $LectureNotesinComputerScience$
32. Yarlagadda, R.T.: Future of robots, ai and automation in the united states. IEJRD-International Multidisciplinary Journal **1**(5), 6 (2015)