Automatic Punctuation Evaluation in Brazilian Educational Texts

Avaliação automática de Pontuação

- Como garantir que a pontuação do aluno está realmente correta?
- Quais as Melhores abordagens?
- Quais as principais dificuldades?

Trabalhos relacionados

- Evaluating performance of grammatical error detection to maximize learning effect
- Intelligent Tutoring System for learning English Punctuation
- The Most Common Punctuation Errors Made by the English and the TEFL Majors at An-Najah National University

Materiais e Métodos

Datasets, modelos e Métricas de Avaliação

- Dois conjuntos de dados usados, um para treinamento, teste e validação e outro para teste.
- 2 modelos usados em duas arquiteturas distintas
- 4 Métricas de Avaliação

Conjunto de Dados

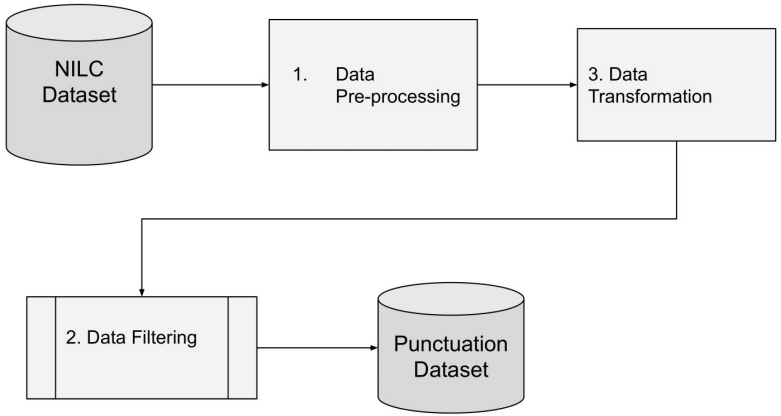
Dataset NILC de livros de diferentes níveis educacionais

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

	Proportion
Ensino Fundamental II	47.7
Ensino Fundamental I	52.3

Conjunto de Dados

Dataset NILC de livros de diferentes níveis educacionais



split	Num. texts	Num Sentences	Sentences Fund. I	Sentences Fund. 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 Dataset

Dataset curado no Projeto Brasil nas Escolas realizados

Labels	Num. Labels
O	33346
I-PERIOD	2190
I-COMMA	896
Total	36432

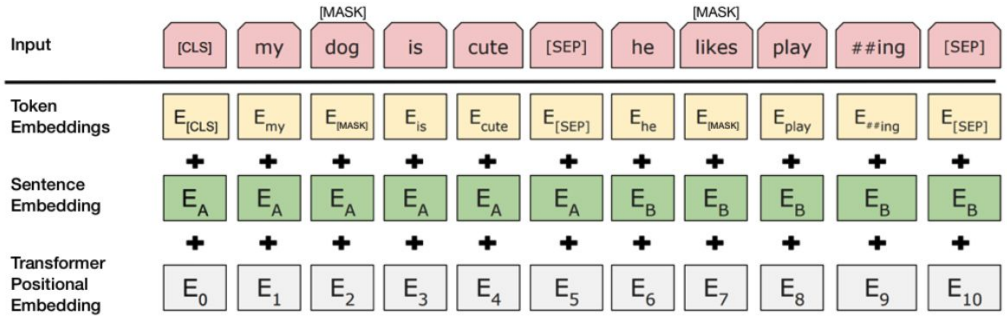
	I-PERIOD	I-COMMA	Total
Inserction	19	80	99
Missing	1205	107	1312
Exchange	49	0	49

Sentences	2190
Texts	265

Errors dos alunos

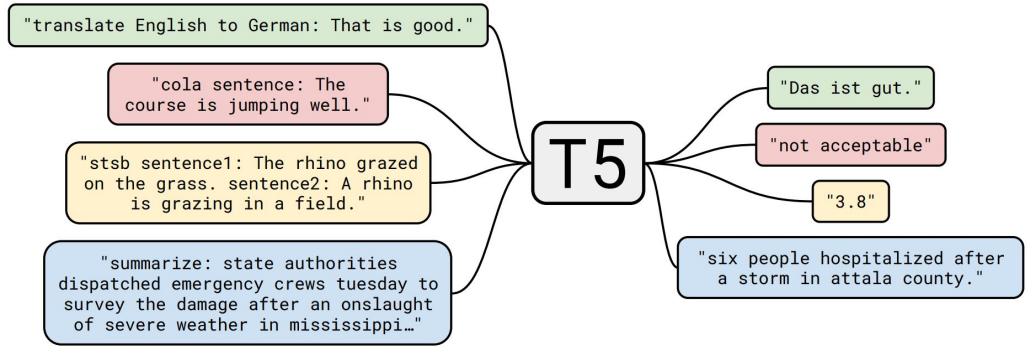
	both annotators	
	I-PERIOD	**I-COMMA**
inserção	19	80
falta	1205	107
troca	49	0

Modelos Usados



[Pre-training of Deep Bidirectional Transformers for Language Understanding]

[Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer]



Arquitetura

BERT			
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
max position embeddings	512	**max position embeddings**	512
type vocab size	2	**type vocab size**	2

T5			
LARGE		**BASE**	
Vocab size	32128	**vocab size**	32128
Dimension model	1024	**d model**	768
Num. of positions	512	**Num. of 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

Evaluation Metrics

BERT

- F1 - score
- Recall
- Precision

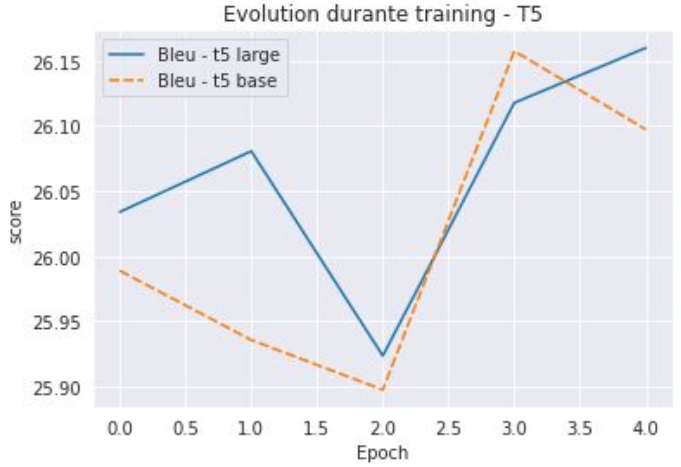
T5

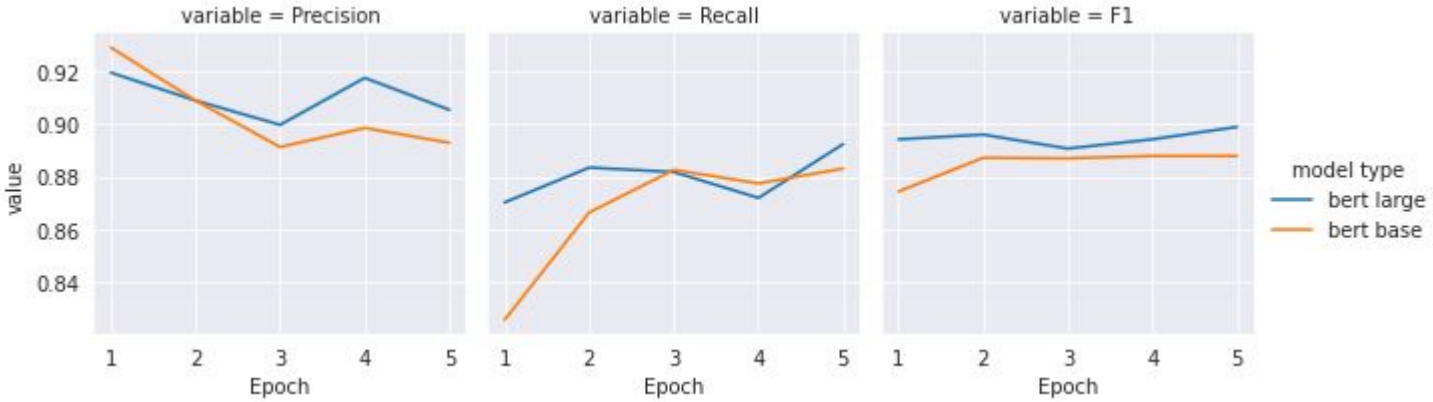
- Bleu score

Training Hyperparameters

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)	Adam with betas=(0.9,0.999)
	and epsilon=1e-08	and epsilon=1e-08
lr scheduler type	linear	linear
num epochs	5	5

Treinamento





Resultados de Teste - Dataset NILC

	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

Resultado MEC dataset

	precision	recall	f1-score	precision	recall	f1-score
	BERT BASE			**BERT LARGE**		
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**		
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

Resultados da Análise

Quantidade total: 438

ID	Caso	Total de Exemplos	Proporção
1	Sentenças com quantidades incorretas de pontuação.	237	54.110
2	Sentenças pontuadas corretamente.	186	42.466
3	Sentenças pontuadas incorretamente e com mesma quantidade de pontuação.	15	3.425

Diferente Número de Labels 237

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**.

Pontuação Correta (186 samples)

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.

Número de Labels Igual, mas mal posicionamento (15 samples)

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**.

Discussão

- O modelo apresentar excelentes resultados para dados do mesmo domínio
- Um possível motivo para a queda significativa nos resultados no outro dataset pode ser causado em parte pela qualidade.
- Realizar o teste em outro dataset poderia ajudar a verificarmos isso.

Conclusão