Punctuation Restoration with Transformers

Krzysztof Wróbel (Enelpol, Jagiellonian University, AGH University of Science and Technology)
Dmytro Zhylko (AGH University of Science and Technology, Enelpol)

Enelpol

Data

Split name	Total size	WikiNews	WikiTalks
rest	22186	13757 (62%)	8429 (38%)
train	800	632 (79%)	168 (21%)
test-A	200	168 (84%)	32 (16%)
test-B	185	185 (100%)	0 (0%)
test-C	274	185 (67.52%)	89 (32.48%)
test-D	244	155 (63.52%)	89 (36.47%)

Example text

rest:

- IN: "To naprawdę wszystko co mogę na ten temat powiedzieć" odpowiedział gdy dziennikarz pytał o bardziej szczegółowe informacje
- EXPECTED: "To naprawdę wszystko, co mogę na ten temat powiedzieć" - odpowiedział, gdy dziennikarz pytał o bardziej szczegółowe informacje.

train:

- IN: " to naprawdę wszystko co mogę na ten temat powiedzieć " odpowiedział gdy dziennikarz pytał o bardziej szczegółowe informacje
- EXPECTED: " to naprawdę wszystko, co mogę na ten temat powiedzieć " odpowiedział, gdy dziennikarz pytał o bardziej szczegółowe informacje.

Preprocessing

- rest dataset
  - Quotation marks and plus characters were separated.
  - Multiple whitespaces deleted.
  - HTML entities were replaced to its characters.

Statistics

	rest		train	
	WikiNews	WikiTalks	WikiNews	WikiTalks
-	6.36%	5.73%	0.00%	0.52%
,	42.75%	41.90%	46.53%	37.86%
;	0.33%	0.58%	0.29%	0.72%
:	3.48%	6.16%	3.55%	5.58%
!	0.25%	1.09%	0.19%	1.43%
?	0.37%	4.49%	0.26%	13.87%
.	46.35%	38.34%	49.18%	39.96%
...	0.12%	1.69%	0.00%	0.06%

Methods

- transformer
- token classification - predict punctuation mark after word

We also experimented with Conditional Random Fields (CRF) layer, but didn't get any tangible improvements.

Experiments

- HerBERT large
- max sequence length: 256
  - text splitted into chunks with overlapping window

1. Few models were trained on *rest* corpus with *train* data as validation
2. Then hyperparameter optimization was performed using *test-A* with *train* corpus as validation over parameters:
   - model: one of few trained on *rest* corpus
   - batch size: 4, 8, 12
   - epochs: 1, 2, 4, 10
   - learning rate: 5e-6, 1e-5, 2e-5, 5e-5
   - evaluation every 50 steps

Results

70 models were trained and the best run achieved the best result on the final leaderboard.

Final model parameters are following:

- batch size 12,
- epochs 10,
- learning rate 5e-6.

Submissions

- AE1 - models trained on *rest*, *train* and *test-A* corpora, without validation, majority voting
- E1 - models trained on *rest*, with validation on *train*, majority voting
- SE1 - further training of models trained on *rest* corpus with *test-A* corpus, with validation on *train*, majority voting
- S1 - further training of models trained on *rest* corpus with *test-A* corpus, with validation on *train*, the best model from 70 models

Results

	test-D weighted F1
AE1	80.09
E1	80.86
SE1	81.27
S1	**81.29**
Norbert Ropiak	81.25
Michał Marcińczuk	81.23

Results of S1 submission per label

Punctuation mark	test-D F1
-	66.91
,	76.08
...	0.00
.	88.68
?	80.61
:	82.98
!	36.36

Error Analysis

```
expected --> predicted
, --> B : 305
. --> B : 171
B --> , : 161
- --> B : 62
. --> , : 47
; --> B : 40
, --> . : 38
: --> B : 35
B --> . : 33
- --> , : 31
B --> - : 31
, --> - : 24
. --> - : 22
- --> . : 19
B --> : : 16
? --> B : 15
? --> . : 14
- --> : : 11
... --> B : 10
. --> ? : 10
. --> : : 10
```

Punctuation errors:

expected: pierwsze zawody, w których wziął udział to niemieckie stanowe mistrzostwa juniorów

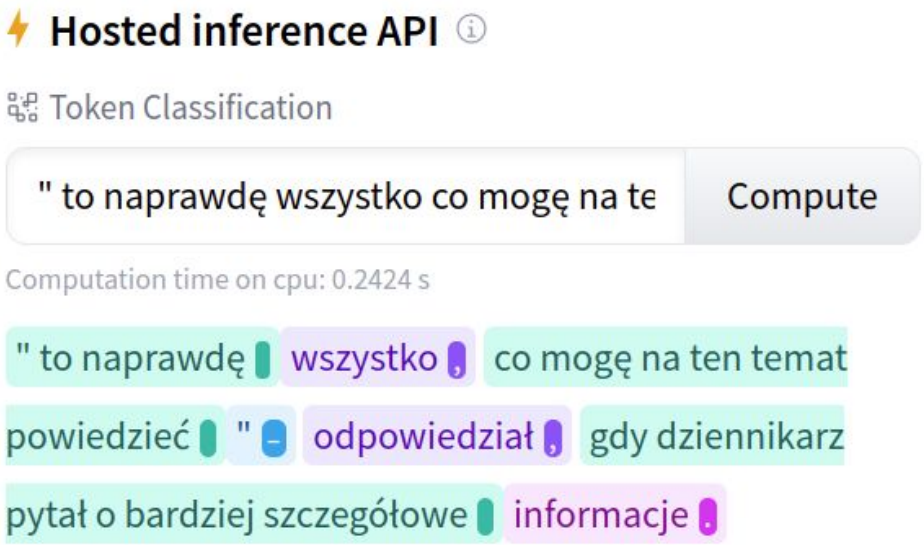
predicted: pierwsze zawody, w których wziął udział, to niemieckie stanowe mistrzostwa juniorów

Also ambiguous punctuations (user decision):

to chyba nie jest właściwe miejsce dla tomografii? - can be both question and exclamation

Hugging Face Hub

https://huggingface.co/enelpol/poleval2021-task1



Source code and models are available at:

https://github.com/enelpol/poleval2021-task1

Contact:

contact@enelpol.com

https://www.linkedin.com/in/wrobelkrzysztof/

https://www.linkedin.com/in/dima-z-a67006131/