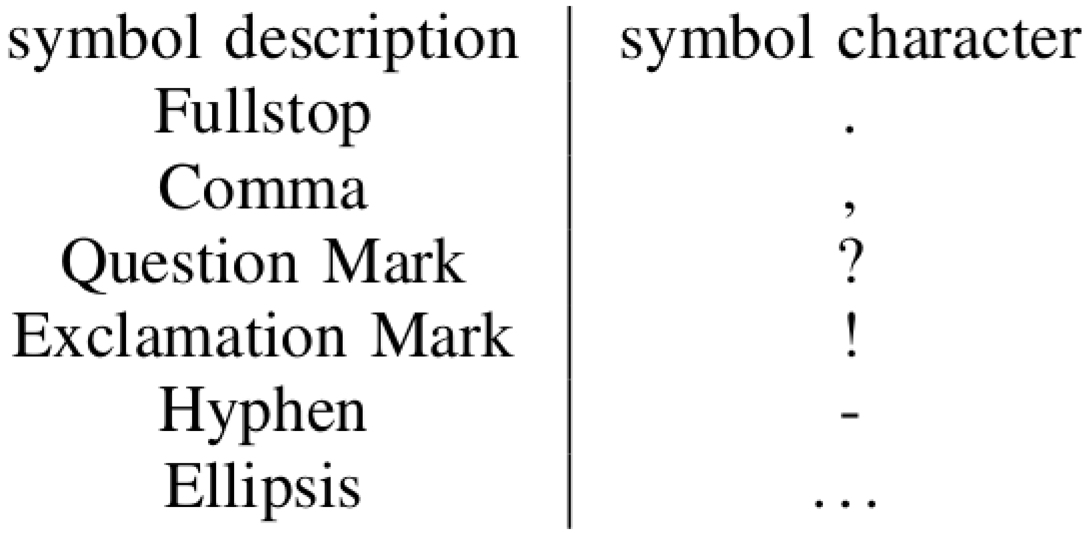
ADAM MICKIEWICZ UNIVERSITY, POZNAŃ

Punctuation Prediction for Polish Texts using Transformers

Jakub Pokrywka

Punctuation Symbols



symbol description	symbol character
Fullstop	.
Comma	,
Question Mark	?
Exclamation Mark	!
Hyphen	-
Ellipsis	. . .

My training data

- Poleval 2022 Task 1: Punctuation Prediction from Conversational Language (this competition training dataset)
- Poleval 2021 Task 1: Punctuation Restoration from Read Text (training dataset)
- Poleval 2021 Task 1: Punctuation Restoration from Read Text (test dataset)
- ~~Europarl-v7.pl-en.pl -only Polish part~~

Sample text

~~**Input wav file** : audio/AU1_P1_w_drodze_do_sklepu.wav~~

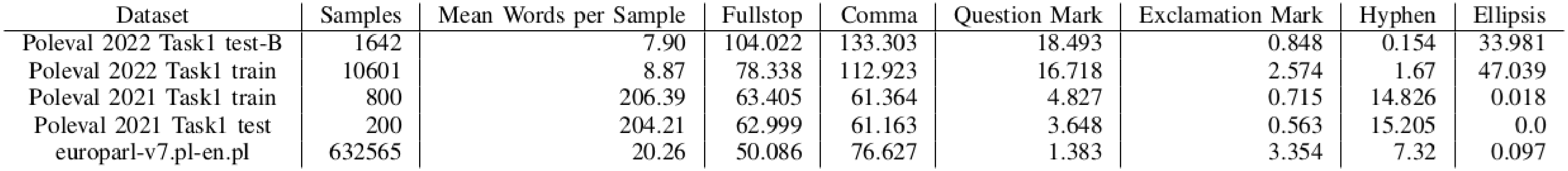
Input text : I:5880-5880 teraz:5940-6180 mamy:6330-6450 drugi:6480-6900 dzień:6960-7080 takiej:7170-7410 ładnej:7440-7650 pogody:7830-8400 Ała:8430-8430 Nie:8760-8820 bij:8850-8970 mnie:9120-9330 kijem:9450-9870 To:10020-10080 boli:10170-10260

Golden truth : I teraz mamy drugi dzień takiej ładnej pogody... Ała! Nie bij mnie kijem! To boli!

Text normalization

- Deleting timestamps and transforming them into regular text
- Replacement of all three consecutive full stop characters "." (Unicode code: 81) with a single ellipsis character ". . . " (Unicode code: 8230) - technical requirement for a utilized library

Punctuation symbols per 1000 words and other statistics



Dataset	Samples	Mean Words per Sample	Fullstop	Comma	Question Mark	Exclamation Mark	Hyphen	Ellipsis
Poleval 2022 Task1 test-B	1642	7.90	104.022	133.303	18.493	0.848	0.154	33.981
Poleval 2022 Task1 train	10601	8.87	78.338	112.923	16.718	2.574	1.67	47.039
Poleval 2021 Task1 train	800	206.39	63.405	61.364	4.827	0.715	14.826	0.018
Poleval 2021 Task1 test	200	204.21	62.999	61.163	3.648	0.563	15.205	0.0
europarl-v7.pl-en.pl	632565	20.26	50.086	76.627	1.383	3.354	7.32	0.097

Utilized python library

FullStop: Multilingual Deep Models for Punctuation Prediction

(SEPP-NLG Shared Task in multilingual sentence segmentation and punctuation prediction)

Utilized included scripts for training as well

Results

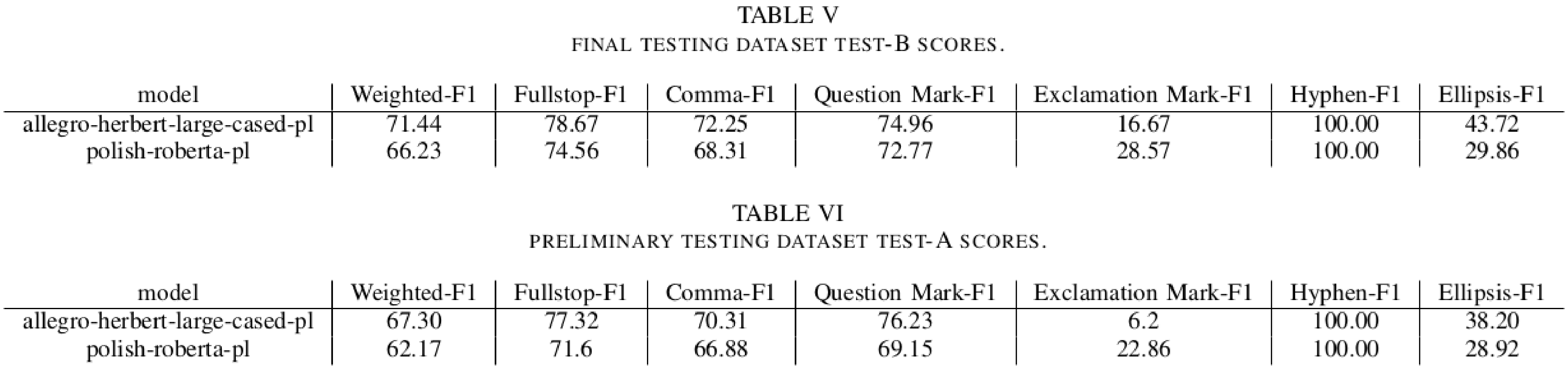


TABLE V

FINAL TESTING DATASET TEST-B SCORES.

model	Weighted-F1	Fullstop-F1	Comma-F1	Question Mark-F1	Exclamation Mark-F1	Hyphen-F1	Ellipsis-F1
allegro-herbert-large-cased-pl	71.44	78.67	72.25	74.96	16.67	100.00	43.72
polish-roberta-pl	66.23	74.56	68.31	72.77	28.57	100.00	29.86

TABLE VI

PRELIMINARY TESTING DATASET TEST-A SCORES.

model	Weighted-F1	Fullstop-F1	Comma-F1	Question Mark-F1	Exclamation Mark-F1	Hyphen-F1	Ellipsis-F1
allegro-herbert-large-cased-pl	67.30	77.32	70.31	76.23	6.2	100.00	38.20
polish-roberta-pl	62.17	71.6	66.88	69.15	22.86	100.00	28.92

Some correct predictions

Predicted: Nie rozumiem powodu, dla którego komuś za ciężko jest rozbić jajko.

Predicted: A ty dasz radę zabrać to wszystko?

Some incorrect predictions

Expected: Ona nie będzie już,
Predicted: Ona nie będzie już…

Expected: Stary d- delegacyjny sprzęt z czasów PRLu, ale może być przydatny.
Predicted: Stary d, delegacyjny sprzęt z czasów PRLu, ale może być przydatny.

Expected: Zamknęli nam łazienkę... dranie...
Predicted: Zamknęli nam łazienkę, dranie

Failed experiments

- Adding word delay
    - **Input text :** I:5880-5880 teraz:5940-6180 mamy:6330-6450 drugi:6480-6900 dzień:6960-7080 takiej:7170-7410 ładnej:7440-7650
    - **Transformed text:** 0 I 60 teraz 150 mamy 30 drugi 60 dzień 90 takiej 30 ładnej
- XLM-RoBERTa

Conclusion

- Fast and relatively easy to implement solution (used ready-made library with training scripts)- but still needs a minor tweak with ellipsis and change in punctuation symbols classes)
- The model training took less than an hour on a single GPU
- The result is much better than the baseline
- But it is significantly worse than two winning solutions

Place	Name	Test-B F1-Score
1	Oskar Bujacz	83.30
2	Michał Pogoda	82.33
3	Jakub Pokrywka	**71.44**
4	baseline	35.30

Thanks for your attention

jakub.pokrywka@amu.edu.pl