How to predict punctuation Quick guide

Oskar Bujacz, University of Wroclaw / QuantUp
FedCSIS 2023, PolEval session

Problem

Goal - prediction of punctuation in Polish language

Source - Speech transcripts generated by Automatic Speech Recognition

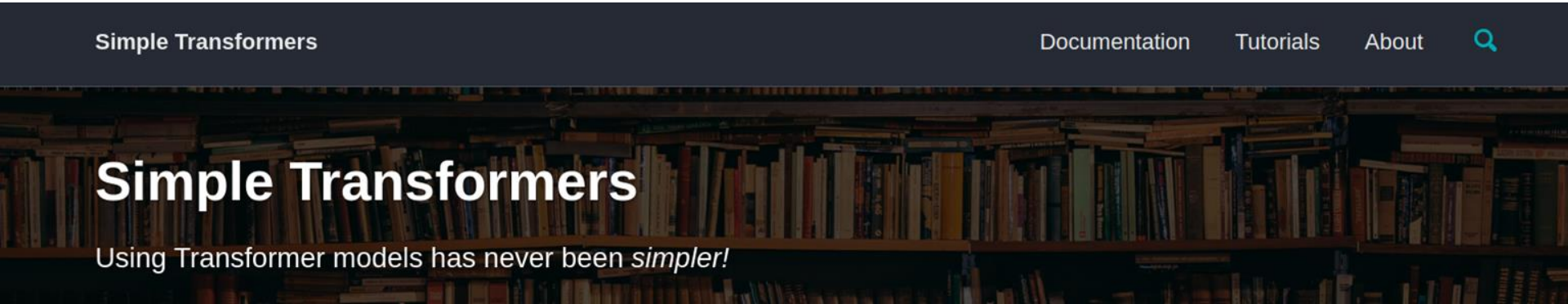
Ambiguous ground truth

Approach

Simplify – treat as a Named Entity Recognition

Example - FEDCSIS conference in Warsaw

Model - allegro/herbert-base-cased with simpletransformers package



Data preprocessing

```
labels = ['O', ',', '.', '-', '...', '?', '!']
```

Problem with ellipsis - unification needed

Merge sentences based on the capitalization

Training details

Focal loss - focus on harder examples

5 epochs, batch size of 20, learning rate 2e-5

Finetuning only on the train dataset

Results!

Task 1: Punctuation prediction

1. Oskar Bujacz (83.30)
2. Michał Pogoda (82.33)
3. Jakub Pokrywka (71.44)

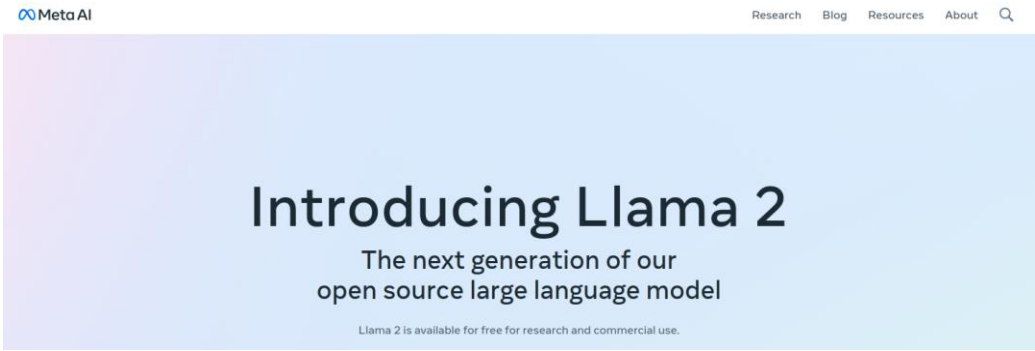
Metric	Score
Weighted-F1	83.30
Hyphens-F1	100.00
Comma-F1	82.83
Ellipsis-F1	60.46
Fullstop-F1	92.59
QMark-F1	80.10
Colon-F1	100.00
Excl-F1	0.00

Possible extensions

More training data

Larger model – LLama2 and its derivatives

Incorporating audio data



Thank you for your attention!