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)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted-F1</td>
<td>83.30</td>
</tr>
<tr>
<td>Hyphens-F1</td>
<td>100.00</td>
</tr>
<tr>
<td>Comma-F1</td>
<td>82.83</td>
</tr>
<tr>
<td>Ellipsis-F1</td>
<td>60.46</td>
</tr>
<tr>
<td>Fullstop-F1</td>
<td>92.59</td>
</tr>
<tr>
<td>QMark-F1</td>
<td>80.10</td>
</tr>
<tr>
<td>Colon-F1</td>
<td>100.00</td>
</tr>
<tr>
<td>Excl-F1</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Possible extensions

More training data

Larger model – LLama2 and its derivatives

Incorporating audio data
Thank you for your attention!