Preproceedings of the of the 18th Conference on Computer Science and Intelligence Systems pp. 1245–1248

Punctuation Prediction for Polish Texts using Transformers

Jakub Pokrywka
Adam Mickiewicz University
Faculty of Mathematics and Computer Science,
Email: jakub.pokrywka@amu.edu.pl

***Abstract*—Speech recognition systems typically output text lacking punctuation. However, punctuation is crucial for written text comprehension. To tackle this problem, Punctuation Prediction models are developed. This paper describes a solution for Poleval 2022 Task 1: Punctuation Prediction for Polish Texts, which scores 71.44 Weighted F1. The method utilizes a single HerBERT model finetuned to the competition data and an external dataset.**

I. Introduction

AUTOMATIC Speech Recognition (ASR) systems produce speech transcripts, which typically do not contain punctuation. This may negatively impact the overall clarity of the transcribed text. For several reasons, punctuation is important:

- Punctuation reduces ambiguity in communication. The Sentences "Let's eat, children" and "Let's eat children" have completely different meanings, but they only vary in a comma.
- Punctuation helps in clarifying the intended meaning of a text. It provides cues to understand the structure of the text. Punctuation marks like commas, periods, question marks, and exclamation marks indicate pauses, sentence endings, and changes in tone or intent.
- Punctuation conveys tone and emotion behind the text. E.g., an exclamation mark may indicate excitement and a question mark may denote uncertainty.
- Punctuation enhances the readability of the written words. Breaking down complex sentences into smaller parts with the use of commas, colons, and semicolons creates pauses, which aids in understanding the text

Many post-processing steps may be taken to circumvent this problem and the lack of capitalization problem. Such tasks are:

- Punctuation Restoration (PR)
- Punctuation Prediction (PP)
- Capitalization Restoration (CR)

The task of Punctuation Restoration is defined as the act of reinstating the original punctuation found in read speech transcripts.

This work describes the solution to Poleval 2022 Task 1: Punctuation Prediction from conversational language. The solution is based on the HerBERT model [1] fine-tuned to the competition data and an external dataset.

II. Related Work

In the previous PolEval edition, a task similar to Punctuation Prediction was assigned, precisely PolEval 2021 Task: Punctuation restoration from read text [2]. The challenge unveiled WikiPunct, a fresh collection of text and audio corpus comprising 39 hours of audio and approximately 38,000 text transcripts. Four submissions [3], [4], [5], [6] applied transformer-based methods for token classification, from which two authors utilized ensembles. Additionally, one author explored the integration of a bi-LSTM layer at the top of the transformer, along with vectors acquired from a wave2vec model.

When it comes down to other languages, authors of [7] developed a method on Support Vector Machines with Conditional Random Field (CRF) classifiers, using part-of-speech (POS) and morphological data for Arabic texts. Authors of [8] used Deep Neural Networks and Convolutional Neural Networks for English texts and authors of [9] used transformers for English medical texts.

Recently, The Sentence End and Punctuation Prediction for many languages shared task was launched [10]. All of the teams explored neural network models, particularly transformers. The winning team described their solution in [11].

III. Competition Description

The three datasets are provided for in the competition: train, dev, and test. For each dataset, input audio WAV files with text transcribed by an ASR system are delivered. The input text is segmented, where a single space separates each word. Each word is prepended by a word start timestamp and word end timestamp in milliseconds.

The missing punctuation symbols are as in table I.

TABLE I
Punctuation symbols in the challenge.

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

The competition dataset is based on three resources summarized in Table II.

Technical track: Challenges for Natural Language Processing (1st Symposium CNLPS 2023)

TABLE II
THE FULL COMPETITION DATASET (TRAIN, DEV, TEST) STATISTICS.

Subset	Corpus	Files	Words	Audio [s]	Speakers	License
CBIZ [12]	DiaBiz	69	36 250	16 916	14	CC-BY-SA-NC-ND
VC	Video conversations	8	44 656	17 123	20	CC-BY-NC
Spokes [13]	Casual conversations	13	42 730	20 583	19	CC-BY-NC

The dataset is split into three subsets as described in Table III.

TABLE III
COMPETITION DATASET STATISTICS SPLIT INTO TRAIN, DEV, TEST.

Dataset	Files	Words	Audio [s]	License
Train	69	98 095	44 030	CC-BY-SA-NC-ND
Dev	11	12 563	4 718	CC-BY-NC
Test	10	12 978	5 874	CC-BY-NC

The annotation scheme is not publicly available during the competition and will described in [14].

There is one sample data from the training dataset in the subsection below.

A. Sample data

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!

B. Utilized Data

In our final solution, we did not use any audio data. Additionally, we decided not to include start and stop timestamps as we did not observe any significant improvement in their score after conducting multiple experiments. Throughout the training process, we experimented with four different sources.

- Poleval 2022 Task 1: Punctuation Prediction from Conversational Language (this competition training dataset)
- Poleval 2021 Task 1: Punctuation Restoration from Read Text [2] (training dataset)
- Poleval 2021 Task 1: Punctuation Restoration from Read Text (test dataset)
- europarl-v7.pl-en.pl [15]

Regrettably, the europarl-v7.pl-en.pl dataset did not lead to a score improvement. Therefore, it was not utilized in our final solution.

We have carried out normalization procedures. Firstly, we transformed the text format from being split with timestamps to raw text format with timestamps included. Secondly, we replaced all three consecutive full stop characters "." (Unicode code: 81) with a single ellipsis character "..." (Unicode code: 8230). This modification was essential for utilizing the punctuation prediction library explained in Section IV.

Table IV presents the statistics for the training datasets used and competition final test data: test-B. Some punctuation marks are more popular than others, which is consequent in all the datasets. There are some differences between training and testing datasets, but they are insignificant. E.g., the Fullstop character is more common in the test-B dataset than in the train dataset (104.022 vs. 78.338). The same stays true for Comma (133.303 vs. 112.923). The PolEval 2022 dataset exhibits much more significant differences than the PolEval 2021 dataset. This is particularly evident in the Mean Words per Sample metric, as well as in most punctuation characters. While some characters like Fullstop, Comma, and Ellipsis are more prevalent in the PolEval 2022 dataset, Hyphen is less frequent, and the Exclamation mark remains relatively unchanged.

Below are samples of golden truths from each dataset, with the last two examples shortened.

1) Sample Poleval 2022 Task 1 test-B sentence: No dzień dobry pani. Tu mi się jakaś opłata za kartę pobrała.

2) Sample Poleval 2022 Task 1 train sentence: I teraz mamy drugi dzień takiej ładnej pogody... Ała! Nie bij mnie kijem! To boli!

3) Sample Poleval 2021 Task1 train sentence: w wywiadzie dla "polski" jarosław kaczyński podkreślił, że informacje dotyczące radosława sikorskiego zagrażają interesowi państwa. "to naprawdę wszystko, co mogę na ten temat powiedzieć "- odpowiedział, gdy dziennikarz pytał o bardziej szczegółowe informacje. premier kaczyński sugeruje, że dobry kandydat po na szefa dyplomacji to np. jacek saryusz- wolski wymieniony polityk zyskał uznanie braci kaczyńskich za dotychczasową działalność w charakterze dyplomaty i dużą wiedzę."

4) Sample Poleval 2021 Task1 test sentence: 801 co znaczy, że beginki "padły ofiarą reformacji"? grzesie2k wpis na słabym poziomie bzdurna informacja o 50 spalonych waldensach; po co w bibliografii pseudonaukowa książka magdaleny ogórek? fragment recenzji z księgarni gandalf: "magdalena ogórek do inkwizycji oraz kościoła ma stosunek jednoznaczny, pisząc o inkwizycyjnej pożodze oraz występkach heretyków spreparowanych przez inkwizytorów, którzy siali spustoszenie oraz o tym jak to w połowie xiii w? duchowni skupiali się na obsadzaniu stanowisk kościelnych, budowaniu zamętu przez interdykty, schizmy i walki, lekceważyli obowiązki duszpasterskie. nie ukrywa też, że jej celem jest próba rehabilitacji heretyków. takie jednoznacznie ideologiczne ustawienie problematyki nie ma wiele wspólnego z prawdą o epoce, obiektywizmem historycznym.

TABLE IV
DATATASETS STATISTICS. THE NUMBER OF PUNCTUATION SYMBOLS IS NORMALIZED PER 1000 WORDS.

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

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

C. Metric

The challenge metric is the Weighted F1 score. The evaluation script is implemented in the GEval evaluation tool [16]. The challenge was hosted on the gonito platform [17]. The final evaluation is done on the test-B dataset on all the domains. The metric definition is meticulously described in Poleval 2021 Task1 summary paper [2].

IV. METHOD

Our method was based on FullStop: Multilingual Deep Models for Punctuation Prediction [11] library. We slightly modified the library to work on a different set of punctuation marks than it was intended to. The final solution model was based on a single HerBERT [1], a neural model of transformer architecture [18] trained on a corpus of Polish texts. The model was finetuned to the data described in Section III-B with the aforementioned text preprocessing steps. We used scripts available at https://github.com/oliverguhr/fullstop-deep-punctuation-prediction/blob/main/other_languages/readme.md. The Polish RoBERTa [19] model was evaluated as well, but not used for the final solution due to worse results. Both evaluations are available in Tables V and VI. We also conducted experiments with XLM-RoBERTa [20], but unfortunately, we did not achieve better results again.

V. RESULTS

The final model using achieved a third-place score of 71.44 in the competition's Weighted F1 category. While it falls behind the first-place score of 83.30 and the second-place score, it still surpasses the baseline score of 35.30. Frequent punctuation symbols like full stops and commas (occurring above ten times per 1000 words) consistently scored between 70 and 80 in F1. However, the F1 scores varied greatly for less frequent symbols, with scores of 16.67, 100.00, and 43.72.

The subsections below illustrate some correct and incorrect predictions from the test-B dataset.

A. Correct predictions

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

Predicted: A ty dasz radę zabrać to wszystko?

B. 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

VI. CONCLUSIONS

In this paper, we proposed our solution to Poleval 2022 Task 1: Punctuation Prediction for Polish Texts. The method uses a single HerBERT model fine-tuned to the competition training data and other external datasets. The achieved score is 71.44, which falls behind the two best solutions but is significantly better than a baseline.

REFERENCES

[1] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik, "HerBERT: Efficiently pretrained transformer-based language model for Polish," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, (Kiyv, Ukraine), pp. 1–10, Association for Computational Linguistics, Apr. 2021.

[2] A. Mikołajczyk, A. Wawrzynski, P. Pezik, M. Adamczyk, A. Kaczmarek, and W. Janowski, "Poleval 2021 task 1: Punctuation restoration from read text," *Proceedings ofthePolEval2021Workshop*, p. 21.

[3] K. Wróbel, "Punctuation restoration with transformers," *Proceedings ofthePolEval2021Workshop*, pp. 33–37.

[4] N. Ropiak, M. Pogoda, J. Radom, K. Gawron, M. Swędrowski, and B. Bojanowski, "Comparison of translation and classification approaches for punctuation recovery," *Proceedings ofthePolEval2021Workshop*, pp. 39–46.

[5] M. Marcińczuk, "Punctuation restoration with ensemble of neural network classifier and pre-trained transformers," *Proceedings ofthePolEval2021Workshop*, pp. 47–53.

[6] T. Ziętkiewicz, "Punctuation restoration from read text with transformer-based tagger," *Proceedings ofthePolEval2021Workshop*, pp. 55–60.

[7] M. Attia, M. Al-Badrashiny, and M. Diab, "Gwu-hasp: Hybrid arabic spelling and punctuation corrector," in *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, pp. 148–154, 2014.

[8] X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 654–658, 2016.

[9] M. Sunkara, S. Ronanki, K. Dixit, S. Bodapati, and K. Kirchhoff, "Robust prediction of punctuation and truecasing for medical asr," *arXiv preprint arXiv:2007.02025*, 2020.

[10] D. Tuggener and A. Aghaebrahimian, "The sentence end and punctuation prediction in nlg text (sepp-nlg) shared task 2021," in *Swiss Text Analytics Conference–SwissText 2021, Online, 14-16 June 2021*, CEUR Workshop Proceedings, 2021.

[11] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," June 2021.

[12] P. Pęzik, G. Krawentek, S. Karasińska, P. Wilk, P. Rybińska, A. Cichosz, A. Peljak-Łapińska, M. Deckert, and M. Adamczyk, "DiaBiz," 2022. CLARIN-PL digital repository.

[13] P. Pęzik, "Spokes- a search and exploration service for conversational corpus data," pp. 99–109, Selected papers from the CLARIN 2014 Conference, 2014.

[14] S. Karasińska, S. Cichosz, and P. Pęzik, "Evaluating punctuation prediction in conversational language," *Forthcoming*.

[15] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X: Papers*, (Phuket, Thailand), pp. 79–86, Sept. 13-15 2005.

[16] F. Graliński, A. Wróblewska, T. Stanisławek, K. Grabowski, and T. Górecki, "GEval: Tool for debugging NLP datasets and models," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, (Florence, Italy), pp. 254–262, Association for Computational Linguistics, Aug. 2019.

[17] F. Graliński, R. Jaworski, Ł. Borchmann, and P. Wierzchoń, "Gonito.net – open platform for research competition, cooperation and reproducibility," in *Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language* (A. Branco, N. Calzolari, and K. Choukri, eds.), pp. 13–20, 2016.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[19] S. Dadas, M. Perełkiewicz, and R. Poświata, "Pre-training polish transformer-based language models at scale," in *Artificial Intelligence and Soft Computing*, pp. 301–314, Springer International Publishing, 2020.

[20] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019.