Punctuation restoration from read text

Elżbieta Jowik, Agata Kaczmarek, Agata Makarewicz

Contents

Introduction to the problem

- The aim of the project was to restore punctuation in the Automatic Speech Recognition of texts read out loud.
- Task was originally introduced on PolEval2021 competition.
- Data source: WikiPunct - text and audio data from Polish Wikipedia.
- Considered punctuation marks:

  - ➢ full stop (.)
  - ➢ comma (,)
  - ➢ question mark (?)
  - ➢ exclamation mark (!)
  - ➢ hyphen (-)
  - ➢ colon (:)
  - ➢ ellipsis (...)
  - ➢ semicolon (;)
  - ➢ quote (“)

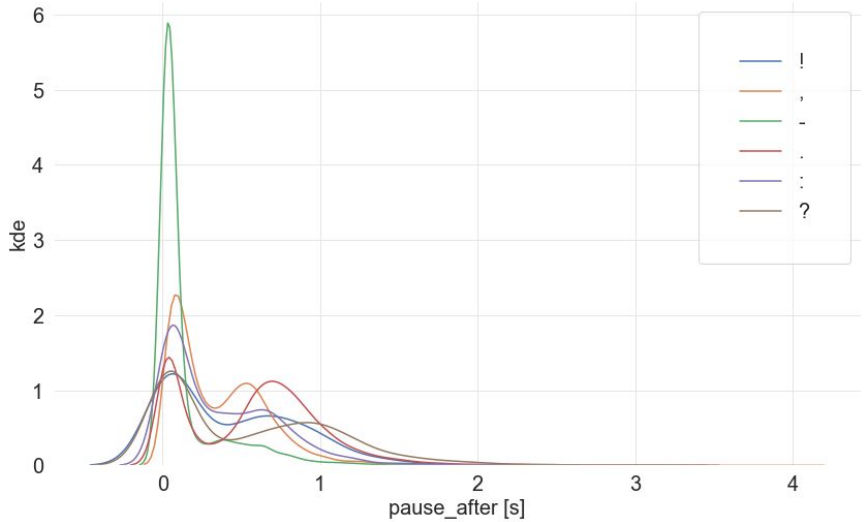
Inspirations from literature

- Morfeusz - Practical Tool for the Morphological Analysis of Polish,
- openSMILE - tool for processing speech,
- Named Entity Recognition Model - token-based unary classifier.

About data

Explanatory Data Analysis

- Text data
  - 800 records
  - Most common punctuation marks: full stop, comma & hyphen.
  - Most frequent words after commas: *że, czy, ale, bo, który, która*.
- Audio data
  - 800 audio, 793 transcriptions for audio
  - Typically bigger pauses correspond to presence of some of the punctuation marks.



Data cleaning & preprocessing

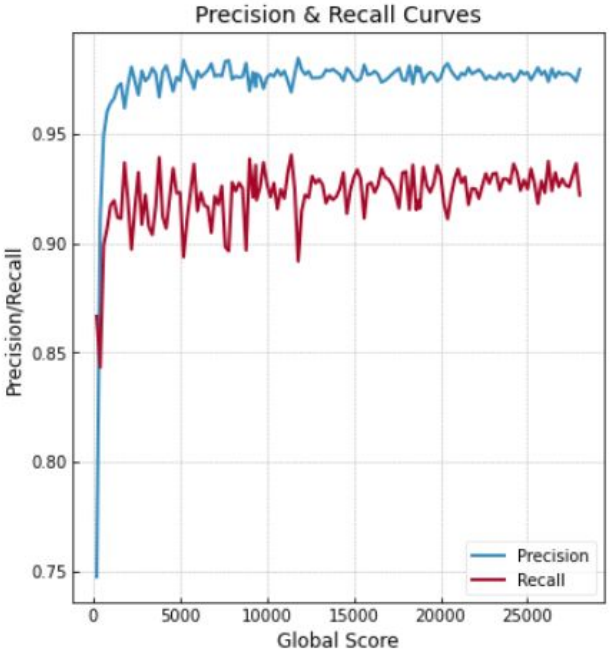
- Deletion of records not having transcripts.
- Use of Morfeusz tool, to find denormalized words.
- Excluding:
  - all words containing non-polish letters
  - all symbols (eg mathematical)
  - punctuation that was not considered during solving the task (bracket, quote, semicolon)

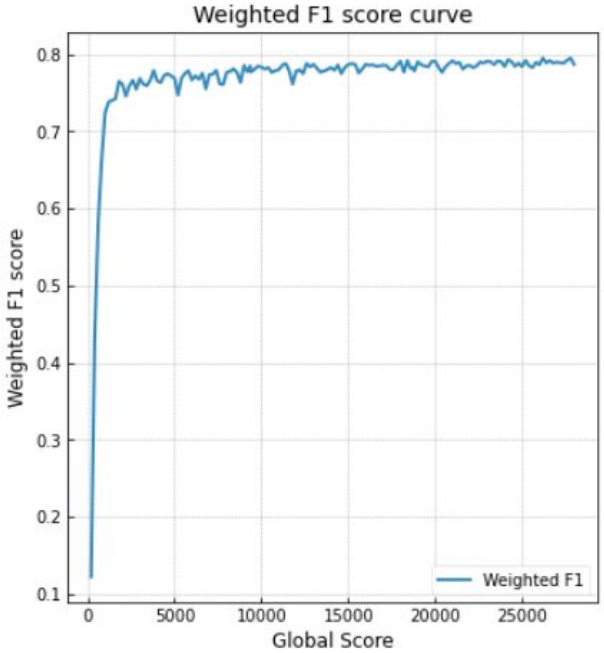
About solution

Solution concept

- Baseline model
  - dictionary-based solution
  - inserting full stops based on pauses
  - inserting full stops in abbreviations
  - inserting commas based on Polish language rules (literature- and experience-based)
- Neural model
  - NER model
  - pre-trained, for Polish language
  - to predict what type of punctuation or none follows a word in a sequence

Results of a NER model





Results of a NER model

Results at the initial stage of learning process

	:	;	,	.	-	?	!
:	0	0	43	131	0	0	0
;	0	0	0	0	0	0	0
,	0	0	787	1543	0	0	0
.	0	0	668	2025	0	0	0
-	0	0	199	412	0	0	0
?	0	0	46	157	0	0	0
!	0	0	9	26	0	0	0

Results at the final stage of learning process

	:	;	,	.	-	?	!
:	621	0	33	95	53	2	0
;	0	0	0	0	0	0	0
,	22	0	8549	523	149	46	17
.	20	0	289	9761	51	38	14
-	45	0	313	273	1520	26	2
?	4	0	28	127	8	539	2
!	2	0	16	49	2	11	22

Overall results

- There is a significant difference in results between two considered approaches for each of the models due to the large majority of blanks among the classes. This class had a great weight assigned and it was also a class recognized best by the models.
- The cardinality of the blank class was the reason of small differences between the results of two baseline models while including blanks, and significantly higher difference while excluding them. In the second approach weight assigned to all the other punctuation marks increases greatly.

	Baseline with pauses	Baseline without pauses	Neural model
Including blanks	0.8248	0.7895	0.9289
Excluding blanks	0.3356	0.1261	0.7229

Conclusions & possible improvements

- Neural model presents good results, however, it has many hyperparameters. It took great amount of time and computational power while training.
- Possible improvements:
  - trying another architecture,
  - more representative dataset - this had only 793 records.

Sources

- http://2021.poleval.pl
- Literature:

  - Devlin, J., Chang, M.W., Lee, K and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018
  - Eyben, F., Wollmer, M. and Schuller, B. "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", 2010
  - Marcińczyk, M. "Punctuation Restoration with Ensemble of Neural Network Classifier and Pre-trained Transformers", 2021
  - Nagy, A., Bial, B. and Acs, J. "Automatic punctuation restoration with BERT models", 2021
  - Ogrodniczuk, M., Kobyliński, Ł. "Proceedings of the PolEval 2021 Workshop", 2021
  - Sandhya, P., Spoorthy, V., Koolagudi, S.G. and Sobhana, N.V. "Spectral Features for Emotional Speaker Recognition", 2020
  - Sopyła, K. "A curated list of Polish abbreviations for NLTK sentence tokenizer based on Wikipedia text", 2020, GitHub Gist
  - Teixeira, J. P., Oliveira, C. and Lopes, C. "Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters", 2013
  - Woliński, M "Morfeusz - a Practical Tool for the Morphological Analysis of Polish", 2006
  - Wróbel, K. and Zhylko, D. "Punctuation Restoration with Transformers", 2021
  - Ziętkiewicz, T. "Punctuation Restoration from Read Text with Transformer-based Tagger, 2021

Thank you for your attention!