Task 1. Poleval 2021

Punctuation restoration from read text

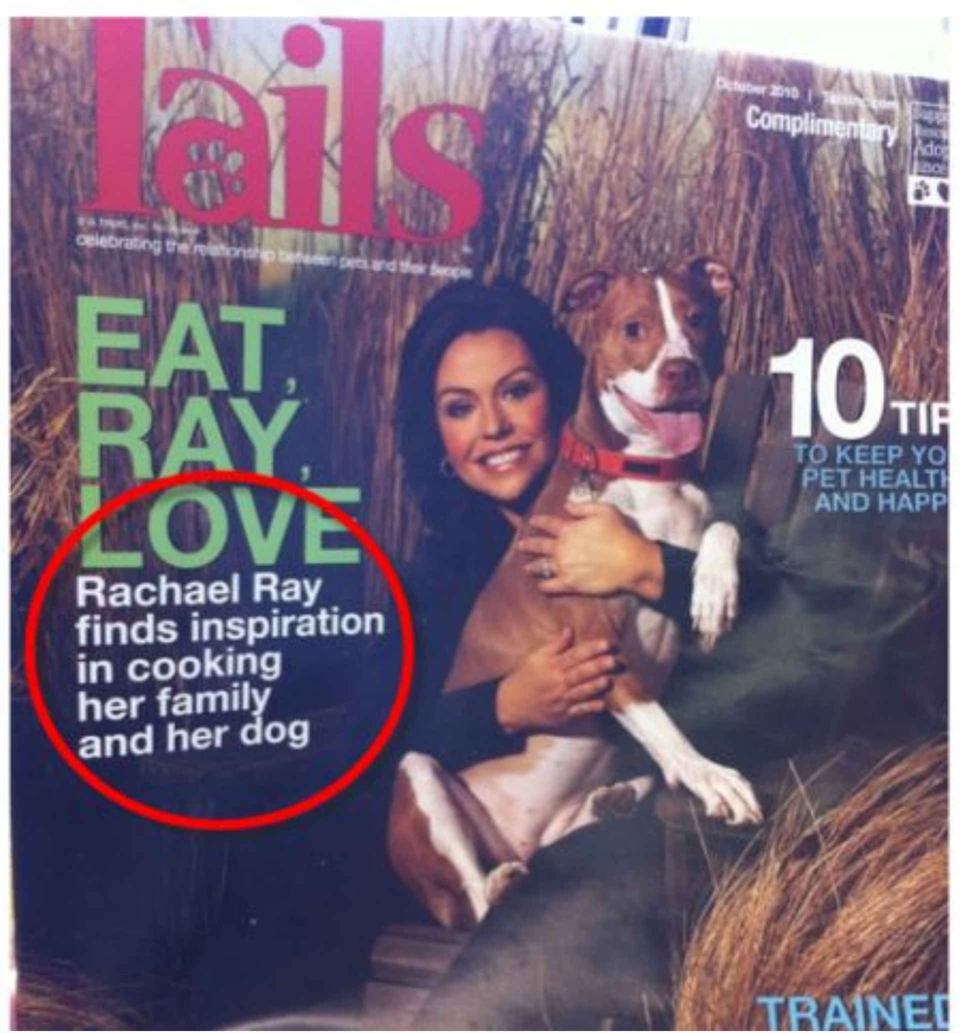
Task authors: Agnieszka Mikołajczyk, Piotr Pęzik, Adam Wawrzyński, Adam Kaczmarek, Wojciech Janowski, Michał Adamczyk

VoiceLab | NLP

Motivation

ASR systems typically provide a transcription without any punctuation

- difficult to read
- poor user experience
- sometimes ambiguous
- old approaches demand pronouncing “full stop”, “comma”, “question mark” out loud - inconvenient

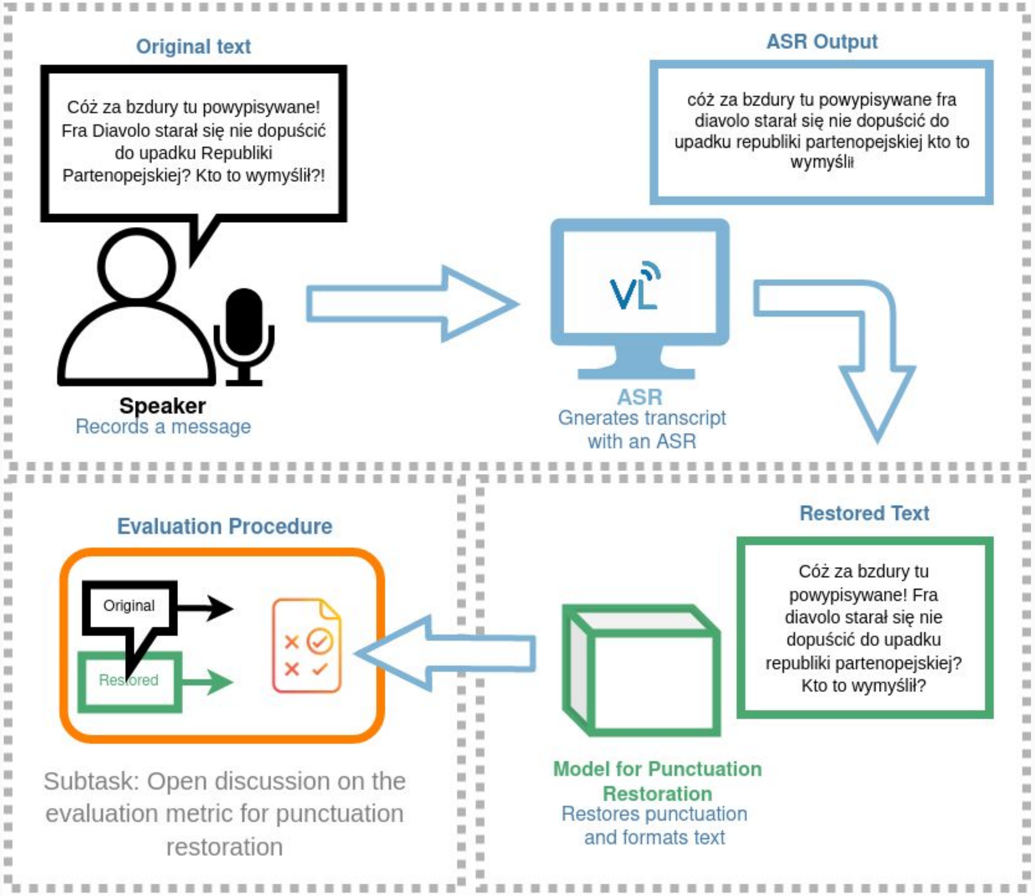

Tails
celebrating the relationship between pets and their people
October 2010
Complimentary
EAT, RAY, LOVE
Rachael Ray finds inspiration in cooking her family and her dog
10 TIP
TO KEEP YO
PET HEALTH
AND HAPP
TRAINED

Motivation

- punctuation improves many NLP downstream tasks e.g.
  - text segmentation
  - named entity recognition,
  - uppercasing
- lack of multimodal punctuation restoration datasets in Polish
- research regarding punctuation restoration in Polish

Task Overview

Dataset WikiPunct



Original text

Cóż za bzdury tu powypisywane! Fra Diavolo starał się nie dopuścić do upadku Republiki Partenopejskiej? Kto to wymyślił?!

ASR Output

cóż za bzdury tu powypisywane fra diavolo starał się nie dopuścić do upadku republiki partenopejskiej kto to wymyślił

Restored Text

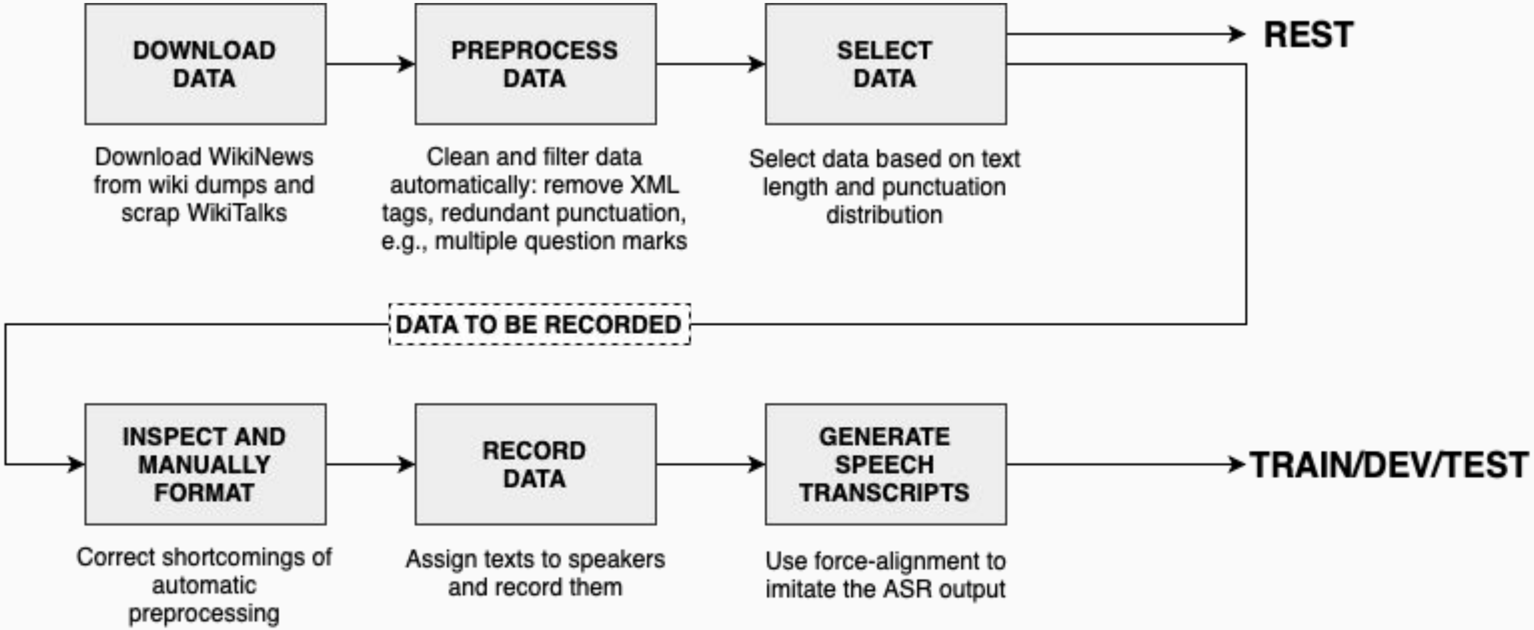
Cóż za bzdury tu powypisywane! Fra diavolo starał się nie dopuścić do upadku republiki partenopejskiej? Kto to wymyślił?

Model for Punctuation Restoration
Restores punctuation and formats text
Evaluation Procedure
Original
Restored

Subtask: Open discussion on the evaluation metric for punctuation restoration

Task. Model for Punctuation Restoration from spoken text.

Data - overview



DOWNLOAD DATA

Download WikiNews from wiki dumps and scrap WikiTalks

PREPROCESS DATA

Clean and filter data automatically: remove XML tags, redundant punctuation, e.g., multiple question marks

SELECT DATA

Select data based on text length and punctuation distribution

REST

DATA TO BE RECORDED

INSPECT AND MANUALLY FORMAT

Correct shortcomings of automatic preprocessing

RECORD DATA

Assign texts to speakers and record them

GENERATE SPEECH TRANSCRIPTS

Use force-alignment to imitate the ASR output

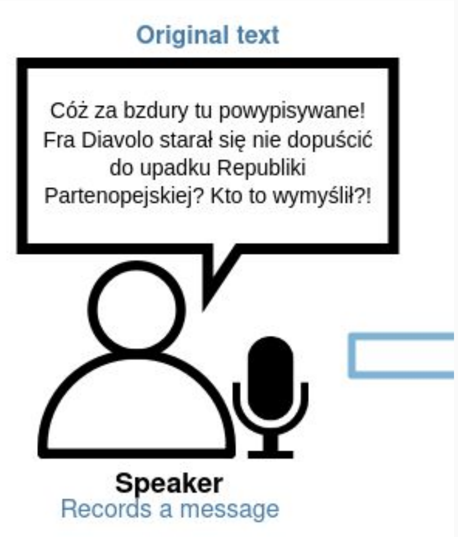
TRAIN/DEV/TEST

Data - overview

WikiPunct

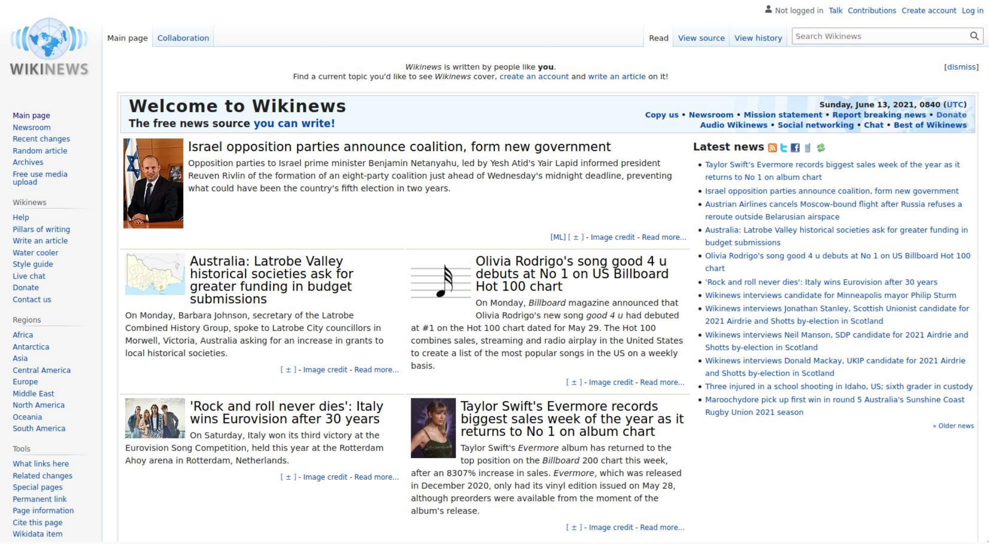
WikiPunct is a crowdsourced text and audio dataset of Polish Wikipedia pages read out loud by Polish lectors.

- Two sources of data:
  - WikiTalks pages - conversational interactive
  - WikiNews - informative
- Original text with punctuation is read by volunteers
- Audio with time-aligned transcripts


Original text
Cóż za bzdury tu powypisywane! Fra Diavolo starał się nie dopuścić do upadku Republiki Partenopejskiej? Kto to wymyślił?!
Speaker
Records a message

WikiNews

- Data scraped from Polish WikiNews via wiki dumps
- WikiNews is a free-content news wiki for collaborative journalism
- Rich punctuation and high overall text quality
- Various subjects and text lengths



Wikinews is written by people like **you**.
Find a current topic you'd like to see *Wikinews* cover, create an account and write an article on it!

Welcome to Wikinews

The free news source you can write!

Sunday, June 13, 2021, 0840 (UTC)

Israel opposition parties announce coalition, form new government

Opposition parties to Israel prime minister Benjamin Netanyahu, led by Yesh Atid's Yair Lapid informed president Reuven Rivlin of the formation of an eight-party coalition just ahead of Wednesday's midnight deadline, preventing what could have been the country's fifth election in two years.

[ML] [±] - Image credit - Read more...

Australia: Latrobe Valley historical societies ask for greater funding in budget submissions

On Monday, Barbara Johnson, secretary of the Latrobe Combined History Group, spoke to Latrobe City councillors in Morwell, Victoria, Australia asking for an increase in grants to local historical societies.

[±] - Image credit - Read more...

'Rock and roll never dies': Italy wins Eurovision after 30 years

On Saturday, Italy won its third victory at the Eurovision Song Competition, held this year at the Rotterdam Ahoy arena in Rotterdam, Netherlands.

[±] - Image credit - Read more...

Olivia Rodrigo's song good 4 u debuts at No 1 on US Billboard Hot 100 chart

On Monday, *Billboard* magazine announced that Olivia Rodrigo's new song *good 4 u* had debuted at #1 on the Hot 100 chart dated for May 29. The Hot 100 combines sales, streaming and radio airplay in the United States to create a list of the most popular songs in the US on a weekly basis.

[±] - Image credit - Read more...

Taylor Swift's Evermore records biggest sales week of the year as it returns to No 1 on album chart

Taylor Swift's *Evermore* album has returned to the top position on the *Billboard* 200 chart this week, after an 8307% increase in sales. *Evermore*, which was released in December 2020, only had its vinyl edition issued on May 28, although preorders were available from the moment of the album's release.

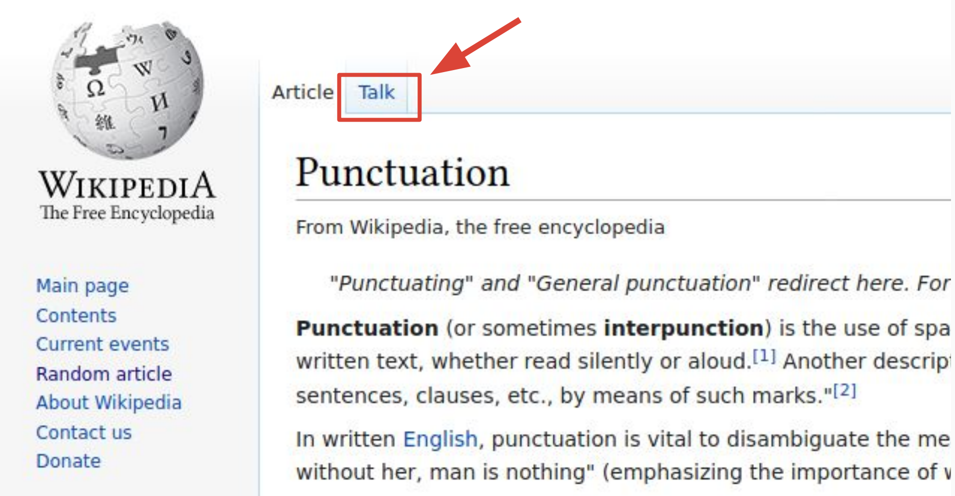
[±] - Image credit - Read more...

Latest news

- Taylor Swift's Evermore records biggest sales week of the year as it returns to No 1 on album chart
- Israel opposition parties announce coalition, form new government
- Austrian Airlines cancels Moscow-bound flight after Russia refuses a reroute outside Belarusian airspace
- Australia: Latrobe Valley historical societies ask for greater funding in budget submissions
- Olivia Rodrigo's song good 4 u debuts at No 1 on US Billboard Hot 100 chart
- 'Rock and roll never dies': Italy wins Eurovision after 30 years
- Wikinews interviews candidate for Minneapolis mayor Philip Sturm
- Wikinews interviews Jonathan Stanley, Scottish Unionist candidate for 2021 Airdrie and Shotts by-election in Scotland
- Wikinews interviews Neil Manson, SDP candidate for 2021 Airdrie and Shotts by-election in Scotland
- Wikinews interviews Donald Mackay, UKIP candidate for 2021 Airdrie and Shotts by-election in Scotland
- Three injured in a school shooting in Idaho, US; sixth grader in custody
- Maroochydore pick up first win in round 5 Australia's Sunshine Coast Rugby Union 2021 season

WikiTalks

- Data scraped from Polish Wikipedia Talk pages
- Talk pages are administration pages with editorial details and discussions for Wikipedia articles
- Good source of conversation-like written data
- Source of questions which are uncommon in WikiNews and Wikipedia Articles
- More punctuation errors than in WikiNews



Punctuation

From Wikipedia, the free encyclopedia

"Punctuating" and "General punctuation" redirect here. For

Punctuation (or sometimes **interpunction**) is the use of spa written text, whether read silently or aloud.[1] Another descript sentences, clauses, etc., by means of such marks."[2]

In written English, punctuation is vital to disambiguate the me without her, man is nothing" (emphasizing the importance of v

Data - recordings

- Part of scraped data was selected to be recorded

Selection procedure:

  - 80% - randomly selected WikiNews (150 < word count < 300 words)
  - 20% - randomly selected WikiTalks (50 < word count < 300 words and at least one question mark)

- Data selected to be recorded was additionally manually corrected
- Speakers were Polish male and female volunteers
- Each speaker read maximum of fifteen texts
- Speakers did not know what is the target task
- Volunteers recorded random texts at home via dedicated Voicelab Platform, with no special equipment

Data - crowdsourcing

- Reading errors
  - skipping fragments of text usually not occuring in conversational speech
    - Examples:
      - dates: “Ocena: Kenraiz 12:36, 7 paź 2008”
      - IP adress: “Zgłasza: 178.43.202.218”
      - nicknames: “--PtrTlr”
  - skipping words from foreign language, e.g. “三尾の大亀; Sanbi no Kyodaigame;”
  - on-the-fly correction of conjugations and inflections during reading text

Data - Statistics

Training data: 1274 recordings with time-aligned transcripts

Test data: at 200 recordings with time-aligned transcripts

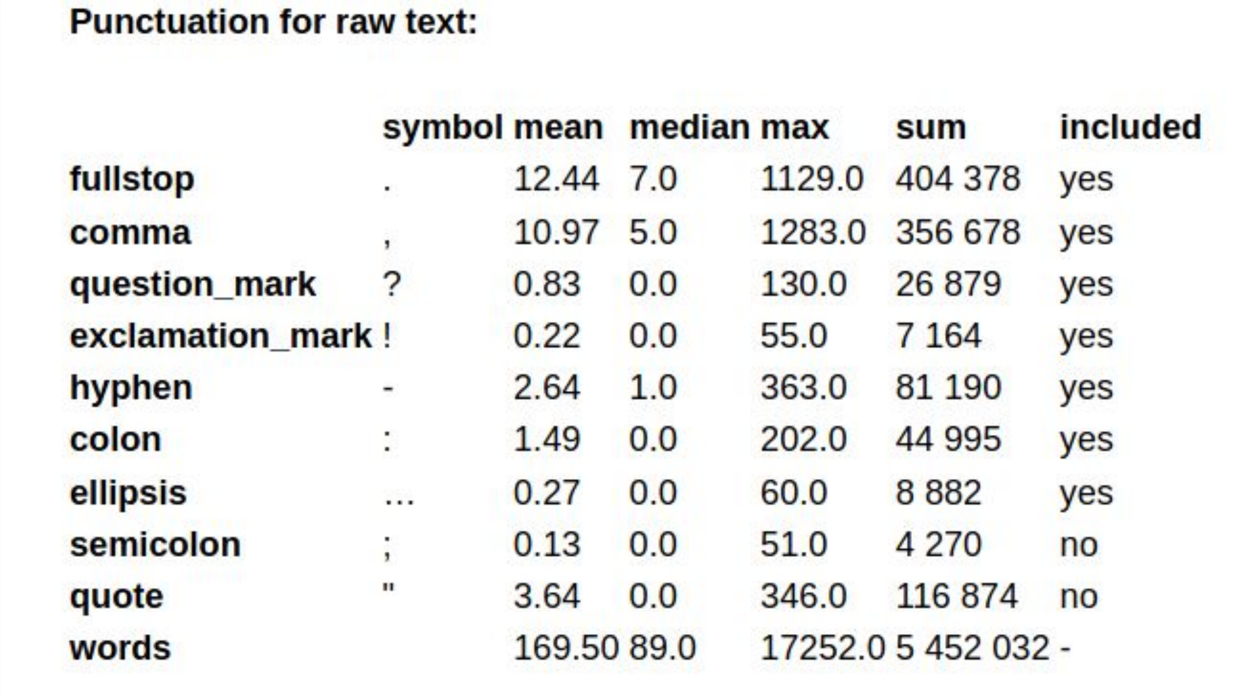
Altogether over 240,000 words

- Speakers:
  - Polish male: 58 speakers, 18.7 hours of speech
  - Polish female: 63 speakers, 20.4 hours of speech

Data - Statistics

Additional text data:

- WikiNews ~15,000
- WikiTalks ~17,000

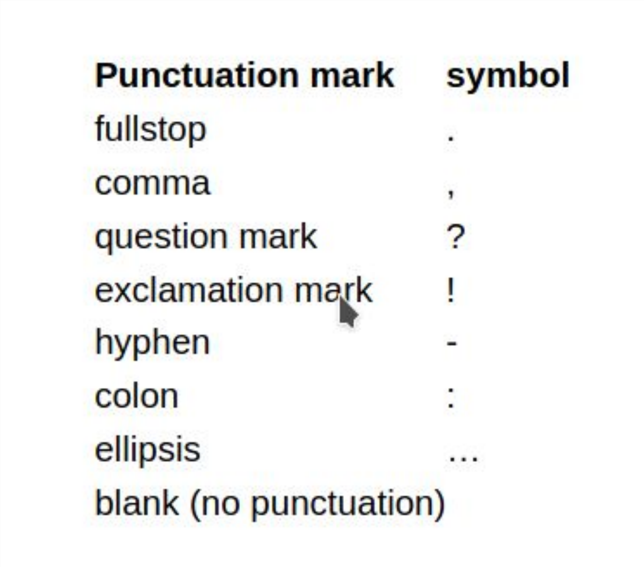


Punctuation for raw text:

	symbol	mean	median	max	sum	included
fullstop	.	12.44	7.0	1129.0	404 378	yes
comma	,	10.97	5.0	1283.0	356 678	yes
question_mark	?	0.83	0.0	130.0	26 879	yes
exclamation_mark	!	0.22	0.0	55.0	7 164	yes
hyphen	-	2.64	1.0	363.0	81 190	yes
colon	:	1.49	0.0	202.0	44 995	yes
ellipsis	…	0.27	0.0	60.0	8 882	yes
semicolon	;	0.13	0.0	51.0	4 270	no
quote	"	3.64	0.0	346.0	116 874	no
words		169.50	89.0	17252.0	5 452 032	-

Evaluation

- Test data: audio with time-aligned transcripts
- Seven punctuation marks are evaluated
- Forbidden to use external data
- Free to use publicly available pretrained models
- Final results are evaluated in terms of precision, recall, and F1 scores for predicting each punctuation mark separately.



Punctuation mark	symbol
fullstop	.
comma	,
question mark	?
exclamation mark	!
hyphen	-
colon	:
ellipsis	...
blank (no punctuation)	

Results

- Token classification
- Bert-like models for feature extraction with FC head
- One approach using wav2vec features from audio

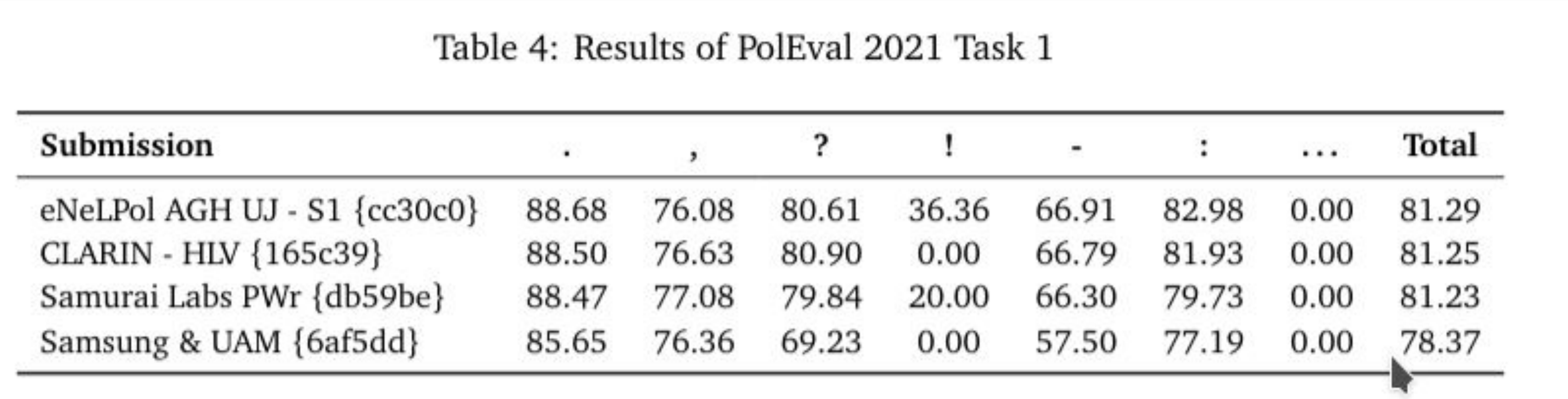


Table 4: Results of PolEval 2021 Task 1

Submission	.	,	?	!	-	:	...	Total
eNeLPol AGH UJ - S1 {cc30c0}	88.68	76.08	80.61	36.36	66.91	82.98	0.00	81.29
CLARIN - HLV {165c39}	88.50	76.63	80.90	0.00	66.79	81.93	0.00	81.25
Samurai Labs PWr {db59be}	88.47	77.08	79.84	20.00	66.30	79.73	0.00	81.23
Samsung & UAM {6af5dd}	85.65	76.36	69.23	0.00	57.50	77.19	0.00	78.37

What’s next?

- Next year we will start new *Task: Punctuation Prediction on Diabiz corpus* together with *Computational and Corpus Linguistics Laboratory at University of Lodz*
- Punctuation prediction instead restoration
- Diabiz is a business conversation corpus
- Diabiz is presented Today by Piotr Pęzik (title: *Budowa referencyjnego korpusu procesów obsługowych i jego zastosowania w tworzeniu systemów dialogowych*)

Any questions?