Samsung Research

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Faculty of Mathematics and Computer Science
Department of Artificial Inteligence

[2]Samsung Poland R&D Institute

Poleval 2021

Task 1: Punctuation restoration from read text

25 października 2021

Tomasz Ziętkiewicz[1,2]

Outline

PolEval

„PolEval is a SemEval-inspired evaluation campaign for natural language processing tools for Polish.”
2021 tasks:

1. Punctuation restoration from read text
2. Evaluation of translation quality assessment metrics
3. Post-correction of OCR results
4. Question answering challenge

Task 1

Goal

“The purpose of this task is to restore punctuation in the ASR recognition of texts read out loud. ”

`http://poleval.pl/tasks/task1/`

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

Problem description

Punctuation is important:

- can change a meaning of a sentence:
  „Powiesić, nie można ułaskawić.”
  „Powiesić nie można, ułaskawić.”

Examples

Twenty five-dollar bills.

$100

Twenty-five dollar bills.

$25

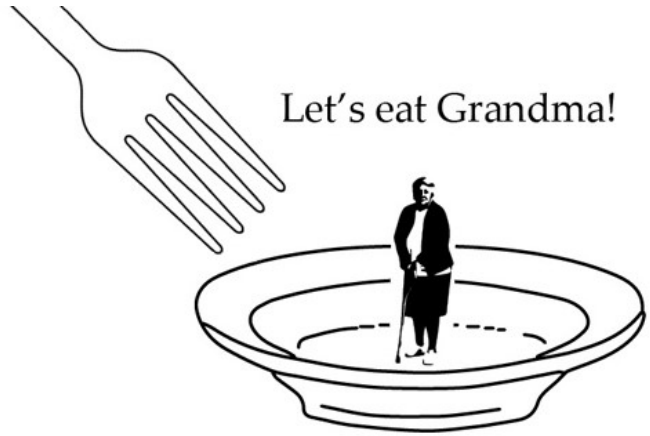
Hyphens matter.

Examples

A woman without her man is nothing.

A woman: without her, man is nothing.

Examples



Problem description

Punctuation is important:

- easier comprehending written text by humans [Moo16].
- people do care about correct normalization
- people prefer a text with correct punctuation [SN21].
- correct punctuation can positively affect NLU performance [Jon95]

Problem description

ASR doesn't care:

- punctuation marks are not directly pronounced
- typically ASR models are trained using reference texts without punctuation marks
- text returned by speech recognition doesn't contain punctuation characters
- need to re-insert them in the post-processing step

Outline

Samsung Research

Data

Dataset origin:

- WikiNews (800 documents)
- WikiTalks (200 documents)

Dataset content:

- Text with punctuation (expected)
- Text without punctuation (input)
- Forced time alignments for each token (all documents recorded and processed by ASR)

No additional data allowed!

Datasets statistics

	Train	Test-A
Documents	800	200
Sentences*	11380	2722
Tokens	165 913	40 842
Max document length (tokens)	313	339
Doc length (sent) (min/mean/max)	2/14/36	3/14/28
!	118	0
?	797	149
:	913	323
-	2448	621
,	10 132	2498
.	10 465	2573

Dataset normalization

- text all-lowercased
- some punctuation characters are tokenized

Samsung Research

Examples

lechia gdańsk pokonała w gdyni arkę 1: 0 w derbach trójmiasta

Examples

(1530,1680) lechia
(1710,2010) gdask
(2040,2400) pokonała
(2550,2550) w
(2670,2910) gdyni
(3060,3630) arkę
(3780,3780) 1:
(3870,3990) 0
(4050,4080) w
(4140,4410) derbach
(4470,4890) trjmiasta,
(5400,5820) rozegranych
(5850,5850) w
(5880,6060) ramach
(6150,6390) smej
(6420,6660) kolejki
(6720,7230) piłkarsk
(7590,8100) ekstraklasy.

Examples

szwa kaszubska w śląskim? a to ci heca! co to w ogóle za mistyfikacja? 21: 19, 3 paź 2010 w etnolekcie ślaskim ” wiadro ” będzie nosiło nazwę ” ajmro ”.

Examples

program rozwoju obszarów wiejskich na lata 2007- 2013 został 24 lipca przyjęty przez komisję europejską

Outline

Samsung Research

Tagging aproach

- Train sequence tagger model on the preprocessed data
- Use the model to tag input text
- Use the tags to restore punctuation

Samsung Research

Data pre-processing

Convert the expected text to text tagged with punctuation characters

- used tokenized version of the expected text (from forced alignments)
- punctuation tokens removed, their labels assigned to preceding tokens
- if there is no preceding token (i.e. the punctuation token is the first in the document) then assign the label to the next token with a special prefix

Example preprocessed data

```
lechia None
 gdask  None
pokonała        None
w       None
gdyni   None
 arkę   None
1       :
0       None
w       None
derbach None
 trjmiasta      ,
```

Implementation

- BertForTokenClassification class from HuggingFace Transformers library - linear layer on top of hidden layer output [WDS+19]
- Hidden layer: HerBERT [MRWG21] pre-trained LM for Polish

Outline

Samsung Research

Evaluation

- Evaluation performed on online competition platform Gonito
- `https://beta.poleval.pl/challenge/punctuation-restoration`
- Results instantly visible to everyone after submission
- Multiple submissions possible
- Evaluation metric: weighted F1-score of each punctuation mark

Results on trainset subset

tag	TP	TN	FP	FN	support	prec.	recall	F1
!	1	16140	0	12	13	1.000	0.077	0.143
,	785	14894	273	201	986	0.742	0.796	0.768
-	135	15862	62	94	229	0.685	0.590	0.634
.	1001	14937	150	65	1066	0.870	0.939	0.903
:	58	16037	29	29	87	0.667	0.667	0.667
?	40	16079	11	23	63	0.784	0.635	0.702
None	13362	2198	246	347	13709	0.982	0.975	0.978
Weighted average with None						0.952	0.952	0.952
Weighted average without None						0.793	0.826	0.806

#	Submitter	Affiliation	Best Submission Description	Task Version	Entries	test-A Weighted-F1 score	test-B Weighted-F1 score	test-C Weighted-F1 score	test-D Weighted-F1 score
1	Krzysztof Wróbel	eNeLPol UJ AGH	S1	3.0.2	9				81.29
2	Norbert Ropiak	CLARIN	HLV	3.0.2	20				81.25
3	Michał Marcińczuk	Samurai Labs, Wrocław University of Science and Technology	qjjdptpc-v4b	3.0.2	21				81.23
4	Dmytro Zhylko	AGH eNeLPol	at least I've tried 🤷	3.0.2	3				80.34
5	Tomasz Ziętkiewicz	Samsung & UAM	Herbert tagger. Spaces after all punctuation marks.	3.0.2	4	78.55	82.32	73.84	78.37
6	Mateusz Piotrowski	None	v2	3.0.2	9	77.42	80.90	72.29	76.65

Next steps

- Joint punctuation and truecasing model
- Models for EU 4 languages
- Experiment with:
- pretrained models used
- data augmentation

Thank you

Thank you for your attention!

Samsung Research

References I

- Bernard E. M. Jones, *Exploring the role of punctuation in parsing natural text*, CoRR **cmp-lg/9505024** (1995).
- Nick Moore, *What's the point? The role of punctuation in realising information structure in written English*, Functional Linguistics **3** (2016), no. 1, 6.
- Robert Mroczkowski, Piotr Rybak, Alina Wr ’oblewska, and Ireneusz Gawlik, *HerBERT: Efficiently pretrained transformer-based language model for Polish*, Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (Kiyv, Ukraine), Association for Computational Linguistics, April 2021, pp. 1–10.

References II

Benjamin Suter and Josef Novak, *Neural Text Denormalization for Speech Transcripts*, Proc. Interspeech 2021, 2021, pp. 981–985.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew, *Huggingface's transformers: State-of-the-art natural language processing*, CoRR **abs/1910.03771** (2019).