OCR Correction with Encoder-Decoder Transformer

Krzysztof Wróbel (Enelpol, Jagiellonian University, AGH University of Science and Technology)

Enelpol

AGH

Data

Dataset	Number of texts	Average number of characters
train	46 097	1378.9
dev	5 508	1372.2
test-A	8 678	1368.6
test-B	8 435	1378.5

Preprocessing

- texts split into chunks because of transformer length limits
- therefore input text and output text alignment
  - token level
  - character level

Alignment example

1. Token level alignment

Input	kroki	nieprzyjacielskle,	*	.		rzekł
Output	kroki			nieprzyjacielskie,“		rzekł

2. Character level alignment

Input	kroki	nieprzyjacielskle,	*	.		rzekł
Output	kroki	nieprzyjacielskie,“				rzekł

Method

Sequence to sequence models:

Model	Parameters [in billions]
plT5-large	0.82
plT5-base	0.28
mT5-base	0.58
mT5-large	1.23
mT5-xxl	13.00

Experiments

- 1 epoch
- batch size: 12
- learning rate: 0.001
- 50-words chunks
- only 100 chunks for validation
- generation with or without sampling

The training time of plT5-large: ~10 hours on GPU Tesla V100.

Results

WER scores:

	test-A	test-B
without correction	18.115	18.114
ED 3 (mT5-base)	4.986	5.001
ED 3 pl (plT5-large)	**4.281**	**4.302**
plT5-base	5.355	5.327
Mateusz Piotrowski (mT5-xxl)	3.725	3.744

Beams

WER scores on full dataset and 100 or 1000 chunks:

Beams	Sampling (100)	Sampling (1000)	No sampling (1000)	No sampling	Sampling
1	5.593	5.711			
2	4.926	5.027	5.073		
4	**4.898**	5.066	4.960	4.992	
8	4.966	**4.911**	4.984	5.003	**4.930**

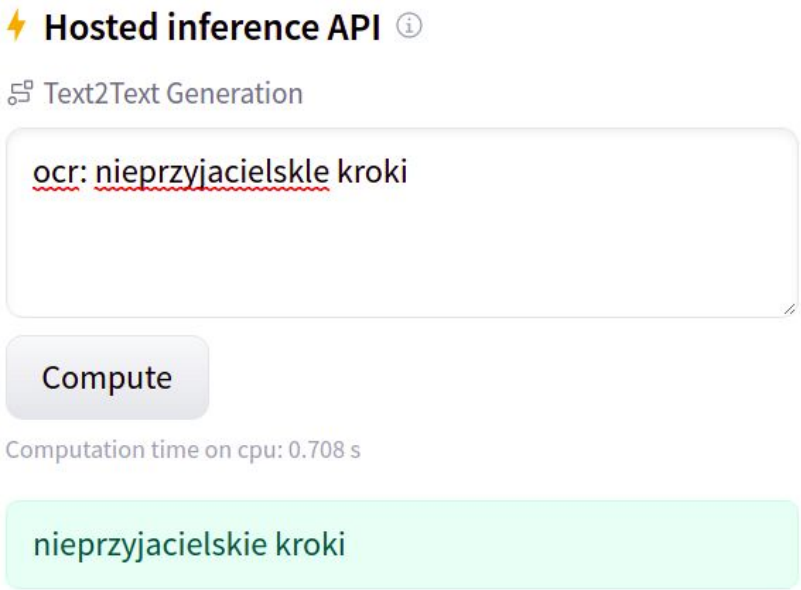
Time of inference

on 100 texts with Tesla V100

	mT5-base		plT5-large		plT5-base	
Beams	No sampling	Sampling	No sampling	Sampling	No sampling	Sampling
1	5.24	10.07	89.88	95.30	16.65	17.94
2	7.59	13.77	139.14	154.91	23.03	27.97
4	10.52	19.32	234.54	263.50	29.16	34.93
8	17.37	34.32	393.75	430.66	41.82	60.19

Hugging Face Hub

https://huggingface.co/enelpol/poleval2021-task3



Source code and models are available at:

https://github.com/enelpol/poleval2021-task3

Contact:

contact@enelpol.com

https://www.linkedin.com/in/wrobelkrzysztof/