PolEval 2021
Task 3: Post-correction of OCR results

Łukasz Kobyliński, Witold Kieraś
Institute of Computer Science PAS
Szymon Rynkun
University of Warsaw

Motivation

OCR methods can only be improved up to a certain level, as the underlying material might be damaged or its image distorted.

Historical texts are often a problem, as they differ from the language expected by a typical OCR method and the quality of print might be low.

Digital libraries provide enormous amounts of scanned documents, which often include OCR layers. The OCR quality is at many times low, but a lot of manual work has been put into filtering out unnecessary elements of such documents and it would be waste to perform new OCR from scratch.

The idea of the task

State of the art deep learning architectures and language models are more and more capable of solving various NLP-related tasks. Let’s evaluate their performance in correcting OCR errors!

Data: collected from Wikisource

Simple data — scanned public domain books, collected from digital libraries and transcribed or corrected by Wikipedia community.

Page-aligned — simulating a real world scenario.

The text is **accurately edited by human editors** including typesetting errors in original transcription.

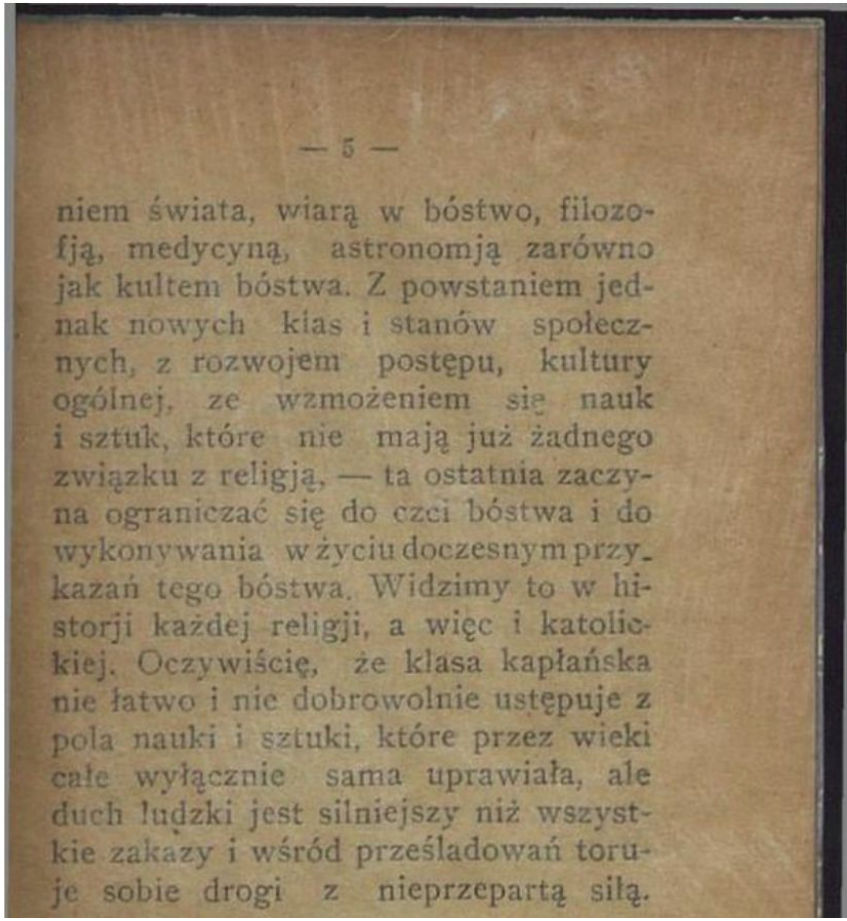
Original text publication date is available and may be used to modify correction algorithm according to a particular chronological period.

Some of the data contains an original **OCR layer**, while we are not using it in this task.

Strona:PL Ignacy Daszyński - Pogadanka o religji.pdf/9

Ta strona została uwierzytelniona.

niem świata, wiarą w bóstwo, filozofją, medycyną, astronomją zarówno jak kultem bóstwa. Z powstaniem jednak nowych klas i stanów społecznych, z rozwojem postępu, kultury ogólnej, ze wzmożeniem się nauk i sztuk, które nie mają już żadnego związku z religją, — ta ostatnia zaczyna ograniczać się do czci bóstwa i do wykonywania w życiu doczesnym przykazań tego bóstwa. Widzimy to w historji każdej religji, a więc i katolickiej. Oczywiście, że klasa kapłańska nie łatwo i nie dobrowolnie ustępuje z pola nauki i sztuki, które przez wieki całe wyłącznie sama uprawiała, ale duch ludzki jest silniejszy niż wszystkie zakazy i wśród prześladowań toruje sobie drogi z nieprzepartą siłą. Prawda nowych nauk przemawiała do umysłów ludzkich z większą potęgą, niż strach klątwy... Kiedy księża kazali słynnemu astronowi Galileuszowi odwołać jego zdanie, jakoby ziemia krążyła około słońca, odwołał on skru-



Preprocessing of the data

The original Wikisource texts contained many inconsistencies.

We removed page numbering from books where it was present (minority of all pages). We have discarded pages, which contained less than 150 characters.

We have discarded “metadata” pages such as “this page contains graphical data”. We have also discarded texts written in **non-Latin scripts** as well as **musical scores**.

We have used **Tesseract** to create the OCR layer of each of the pages.

A few words about OCR methods

While we originally planned to use several OCR methods, we had to settle with a single algorithm because of licensing issues.

We have finally used tesseract-ocr for all books (Apache licence, version 4.0.0-beta.1).

We have used Tesseract with all settings left at default values.

Preprocessing of the data (cont.)

As most of the books were **published in the 1930s**, we have reduced the number of data from this period.

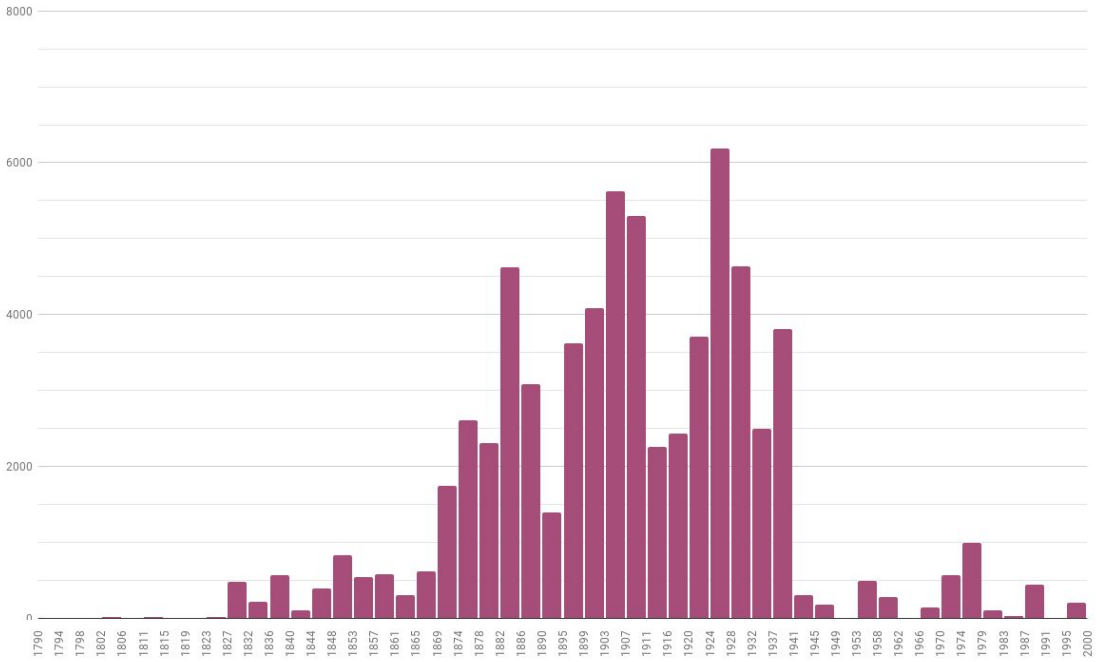
We have discarded **title pages, table of content pages, advertisements, image captions** etc. as such pages collided with the main idea of correcting the running text of books.

Some statistical data

The final dataset contained:

- 979 books
- 68 718 pages

Ca. 46 000 pages included in training data, 5 500 pages in development data.



Number of pages with regard to publication date.

The input data format

in.tsv

```
745 26  1907    ca Wtem\n\nwśród postów i ciężkich umartwień,
nad\ngrobem własną ręką wykopanym, w nieu-\nstannem rozmyślaniu o
śmierci [...]
```

1. Document ID
2. Page number
3. Year of publication (not always available)
4. OCR text (input for correction)

Evaluation method

We use word-error rate (WER) on aligned data to calculate final scores.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where

- S is the number of changed tokens,
- D is the number of deleted tokens,
- I is the number of inserted tokens,
- C is the number of correct tokens,
- N is the number of tokens in gold-standard data (N=S+D+C)

Specific implementation used during the task:
https://gitlab.com/filipg/geval

Results

There were **28 submissions in total** for the OCR correction task.

Five teams decided to submit their solutions for the final test-B dataset.

The differences in performance of the systems were substantial and ranged from **WER>8 to WER=3.744** for the final test-B dataset.

The approaches to the task included using **heuristics, custom transformer architectures and a pretrained mT5 model**.

Results

Affiliation	System name	test-A WER	test-B WER
None	XXLv1	3.725	3.744
eNeLPol UJ AGH	ED 3 pl		4.302
Adam Mickiewicz University & Applica.ai	uml	5.338	5.328
None	1		7.208
Samurai Labs, Wrocław University of Science and Technology	Heuristics	8.063	8.217

Conclusions and Future Work

Ideas for future editions of the task:

- using several different OCR methods (e.g. FineReader, Tesseract, in equal proportions),
- including different document types apart from printed books (e.g. typescript, library index cards, severely damaged sources),
- separating the error detection task from error correction task, expecting the algorithm to suggest more than one corrected token instead of a single proposition.

Thank you!