January 2023

CCS HTR Engine Benchmark

Introduction

In over 20 years of experience in content conversion, CCS has collected many datapoints on performance of various OCR engines. From time to time, we conduct evaluations and provide them to our partners. After integrating HTR into our product docWizz in 2021, we have endeavored to build a set of ground truth for training and evaluating HTR systems. In this whitepaper we share the results of our first evaluation conducted.

OCR engines evaluated

We have chosen four systems to evaluate:

Tesseract version 4, a free Open Source (OS) system originally developed by Hewlett-Packard, then for some years sponsored by Google.

Calamari, a free OS system derived from OCRopy and Kraken.

Transkribus, a commercial offering by READ Corporation, Austria. The system was originally developed in the EU founded projects tranScriptorium and READ (Recognition and Enrichment of Archival Documents)

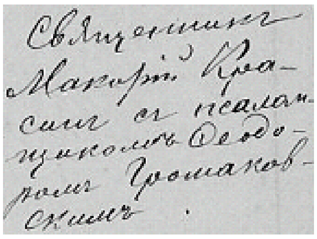
Glyph, a still experimental closed source system developed by a group of researchers at Polytechnic University of Bucharest.

Aim and method of evaluation

We aim to evaluate HTR Engines as a technology. We do not evaluate HTR Services. Therefore, OCR services by Google, Microsoft or Amazon are not included in the list above. They are certainly great services, but the technology is tied with the models and any assessment depends strongly nature of the test data used. In our projects we are often faced with old handwriting in less common languages that require training of specific models. Thus, we are more interested in the technical capabilities of engines to train custom models.

Ground Truth

Our set of ground truth comprises of 590k words in Ukrainian (Cyrillic) language and was collected in a single project. The original documents are from 18th and 19th century. Training of models was started "from scratch", no pre-trained models were used for re-training. We used 98% of our data for training. A 2% random set of the data was excluded from training and used for evaluation. No post-processing like dictionary checks were used in this evaluation.



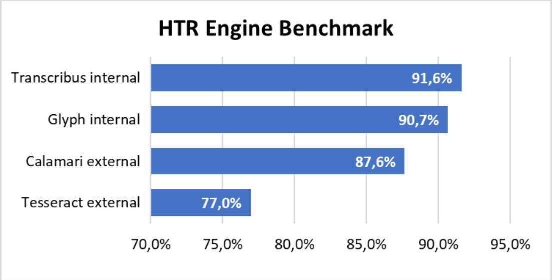
Свидетели
Макарій Кра-
сии и пелан-
цыковъ Федо-
ровъ Громаков-
скимъ .

Line Segmentation

HTR engines usually segment images into lines first. Training and recognition are based on single lines. In our experience, the quality of the line segmentation can have a strong impact on the recognition rates. To understand this effect, we did evaluations twice, once with the internal line segmentation provided by the HTR engine and again with an external line segmentation framework. These evaluations include the training of the models.

In the external case, the engines only "see" images with one line of text. For external line segmentation, we used the OCR-D framework, an OS framework funded by the "Deutsche Forschungsgemeinschaft". With Transkribus, external line segmentation was not used for technical reasons. Calamari does not have an internal line segmentation. Glyph performed slightly better with its internal segmentation. Tesseract improved significantly from external line segmentation.

Results



We use the Levenshtein Distance as metric to count errors. The percentage values are calculated based on characters (including blanks).

Conclusion

Overall, we are positively surprised that a less than 10% error rate was achieved by two engines without postprocessing. Transkribus gives the best results. It not only gives the best results but has proven to be stable and robust against changes in style of writing in other contexts. Glyph performs surprisingly well, especially in consideration of its minimal development budget. Calamari as a free Engine performs quite well but falls short of Transkribus. Tesseract disappoints and seems not to currently be an option for handwriting even though it gives very good results for printed material.

Our data has a focus on names, numbers, and places. Even though we have not seen indicators of it, there may be a risk that this characteristic of the training data impacts the relative performance of the engines.