October 2020

CCS OCR Engine Bechmark

Introduction

In over 20 years of experience in content conversion, CCS has collected many datapoints on performance of various OCR engines. Only recently have we endeavored to build our own tool to enable an automated evaluation of new OCR engines and versions, and the systematic storing of ground truth data. While we still consider our repository of ground truth quite small, we believe that the results are of interest to our partners.

Ground truth data

For the results presented here, we used a collection of ≈80 documents that were passed through layout analysis in our software, docWorks. Random text zones were transcribed using the double keying method. In total, our data adds up to ≈100k characters in over ten different languages. The majority of the data is in English, German and recently also Arabic. The Latin alphabet data is split equally between Antiqua and Fraktur. 80% is newspaper material, the rest are monographs.

Method of evaluation

We use the Levenshtein Distance as metric to count errors. The percentage values are calculated based on characters (including blanks), and special characters are standardized to enable the distance to be computed, e.g. ſ=s.

OCR systems evaluated

For the results presented here, we used ABBYY FR11 and FR12, each in normal (N), balanced (B), and fast mode, Google Vison API (GV), Kofax OmniPage 18 and 20, as well as Tesseract 3.05. Additionally, Tesseract 4.1 (T4.1) was evaluated with the trained models provided as part of the standard distribution, with the exception of Fraktur, where we used our own models.

Results

Fig 1 shows the results of the top three performing engines for each of the three font types Antiqua, Fraktur and Arabic. OmniPage has not made it into the top three of either font types.

It is interesting to note that FR still achieves consistent good results across the board, making it a sort of Swiss Army knife multitool application in terms of OCR - however it only comes first once, for Antiqua. In second place after ABBYY we have Tesseract 4.1, with 97.6% accuracy for Antiqua.

For Fraktur, our choice of trained models for T4.1 clearly outperform FR12. Most surprisingly, GV is in the lead for Arabic, with excellent results. While our data indicates that GV outperforms FR12, the absolute error rate looks almost too good to be true. We assume this stems from the very good quality of images that we used for Arabic, whereas for Antiqua a broader range of qualities and multiple languages was used.

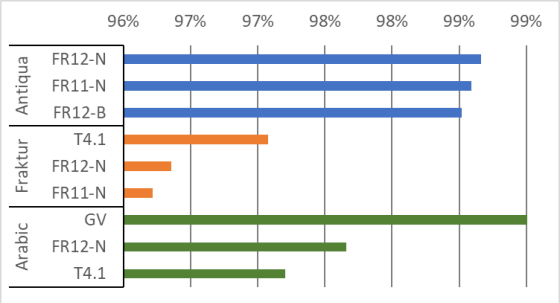

Figure 1 Top 3 OCR systems per font type

Diving a bit deeper into the results on Fraktur, we see in Fig. 2 that our models for T4.1 outperform FR12 in all four languages we had sufficient data for. The relatively poor performance for Danish and Finnish is probably down to our having a limited amount of training data covering the special characters involved.

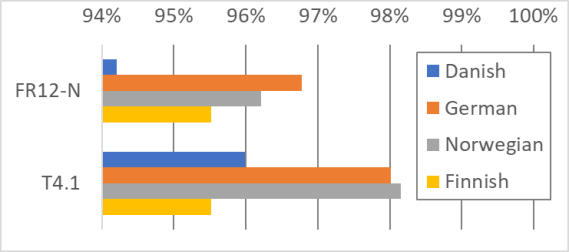

Figure 2 Fraktur results per language

Conclusion

Our results should certainly not be used as sole criterion for the choice of a suitable OCR system for any specific project. However, as a guideline, we recommend that ABBYY, Tesseract and Goggle Vision be included in any selection process. We are pleasantly surprised by the Google Vison results on Arabic material and will strive to extend our ground truth database for further evaluations. Encouraged by the excellent results on Fraktur with our own model for Tesseract 4.1, we intend to grow our repository of that set of training data as well.