Optical Character Recognition with a Neural Network Model for Printed Coptic Texts

Kirill Bulert
kirill.bulert@stud.uni-goettingen.de
eTRAP Research Group
University of Göttingen, Germany

So Miyagawa
so.miyagawa@mail.uni-goettingen.de
University of Göttingen, Germany

Marco Büchler
mbuechler@etrap.eu
eTRAP Research Group
University of Göttingen, Germany

Introduction

Optical character recognition (OCR) is the process of extracting text from images. The final results are machine readable versions of the original images. Nowadays every modern scanner comes with some kind of OCR, but the results may not be satisfying when the OCR is applied to historical texts, that

1. do not use standard fonts,
2. are not printed by a machine,
3. have varying paper and font quality.

Furthermore, historical texts are not passed down through the centuries in their entirety but rather contain lacunae and fragmentary words. This makes automatic post-correction more difficult on historical texts than on modern ones.

We used two tools to create language- and even document- specific recognition patterns (or so-called models) to recognize printed Coptic texts. Coptic is the last stage of the pre-Arabic, indigenous Egyptian language. It was used to create a rich and unique body of literature: monastic, "Gnostic," Manichaean, magical and medical texts, hagiographies, and biblical and patristic translations. We found that Coptic texts have properties which make them excellent candidates for reading by computers. The characters can easily be distinguished due to their limited number and the fact that almost all the hand-written texts exhibit characters with highly consistent forms.

Related Work

The process of digitizing historical documents can be split up into at least three major steps: (1) pre-processing, (2) text prediction (OCR), and (3) post-processing or correction.

Although many works already tackled subproblems (He et al, 2005; Gupta et al, 2007; Kluzner et al, 2009), Springman et al.(2014) presented the first complete approach containing all major steps for historical Greek and Latin books.

The first OCR results for printed Coptic texts were achieved by Mekhaiel (see Moheb's Coptic Pages) by using Tesseract to create a model for Coptic texts. Tesseract assumes that the image was printed with a standardized font. Although it can be trained to use many different fonts, creating a general model that would satisfy scholars is not feasible. In the end, this model is sufficient for pure printed Coptic texts, but creates a lot of noise for texts with mixed languages or annotations. Such drawbacks can be easily overcome by checking against a dictionary, but historical languages often do not have a dictionary that could be considered complete, and the texts might only be fragments that require further analysis.

The recognition itself is performed by either Ocropy (Breuel, 2008) or Tesseract. Potentially, all character-based texts can be recognized. However, even though Mekhaiel provided a Coptic model for Tesseract, we were never able to achieve satisfying results on images which were not pre-processed.

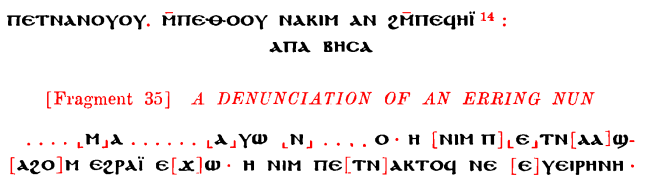
Data Used

For training and testing, an expert on Coptic created a clean version and transcription of Kuhn's 1956 edition "Letters and sermons of Besa." This will also be made available to the interested public.

Besa is a fifth-century abbot of a monastery in Upper Egypt and Coptic writer, whose literary legacy consists mainly of letters to monks and nuns on questions of monastic life and discipline.

Simplified pages were created to find the limits of the trained models with optimal input data. Since creating simplified pages consumes a lot of time, we consider this task as impractical for real use scenarios. Nevertheless, the results on these simplified pages show the best possible prediction.

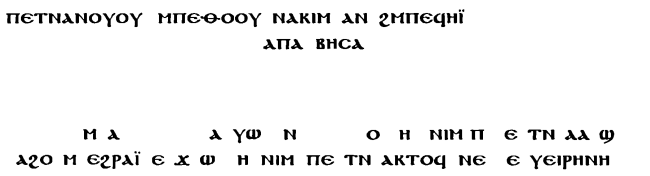
In Fig. 1 all characters and symbols that are going to be removed are marked red. The resulting simplified image can be seen in Fig. 2. By procedure, adjacent characters that are supposed to form one word are cut apart by gaps. Those gaps are going to be predicted differently by the two OCR engines.


ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ⲥⲙ̄ⲡⲉⲫⲏⲓ̈ [14] :
ⲁⲡⲁ ⲃⲏⲥⲁ

[Fragment 35] *A DENUNCIATION OF AN ERRING NUN*

. . . . ⸢ⲙ⸣ⲁ . . . . . ⸢ⲁ⸣ⲩⲱ ⸢ⲛ⸣ . . . . ⲟ · ⲏ [ⲛⲓⲙ ⲡ]⸢ⲉ⸣ⲧⲛ[ⲁⲁ]ⲱ-
[ⲁⲍⲟ]ⲙ ⲉⲍⲣⲁⲓ ⲉ[ⲭ]ⲱ · ⲏ ⲛⲓⲙ ⲡⲉ[ⲧⲛ]ⲁⲕⲧⲟⲩ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

Fig.1, Original Image (excerpt), red elements are missing in the simplified version


ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ ⲙⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ⲥⲙⲡⲉⲫⲏⲓ̈
ⲁⲡⲁ ⲃⲏⲥⲁ

ⲙ ⲁ ⲁ ⲩⲱ ⲛ ⲟ ⲏ ⲛⲓⲙ ⲡ ⲉ ⲧⲛ ⲁⲁ ⲱ
ⲁⲍⲟ ⲙ ⲉⲍⲣⲁⲓ ⲉ ⲭ ⲱ ⲏ ⲛⲓⲙ ⲡⲉ ⲧⲛ ⲁⲕⲧⲟⲩ ⲛⲉ ⲉ ⲩⲉⲓⲣⲏⲛⲏ

Fig. 2, Simplified version (excerpt)

Methodology

There are two methods to train for Coptic texts:

(i) Tesseract needs a font as the baseline and matches the found letters against this font. This can be highly convenient since fonts do not show many variations within a single document. Additional fonts can be incorporated into the model with the drawback that the prediction requires more computational time. So far, we have used Mekhaiel's original model, and we are currently experimenting by adding document-specific characters to increase the accuracy of a single document.

(ii) Ocropy, on the other hand, does not require a font. For training, it requires only a partial transcription: the ground truth. This transcription is used to train a neural network that can recognize the characters. Ocropy's drawback is that the ground truth cannot just be the alphabet but requires multiple pages of transcribed text with a representative letter frequency. Ocropy's training process is measured in iterations. Springmann proposed working with at least 30,000 iterations (a comment made by Springmann in a private conversation, based upon his own experience).

For this contribution, we created an Ocropy model with a training set containing approximately 5,000 characters. This set includes superlinear strokes, braces and foreign characters which are not part of the Coptic alphabet.

Multilingual documents and documents containing foreign characters are considered complex. Stains on the document, bad image quality, and annotations like line numbers increase the complexity of documents as well. We, therefore, created special pages with reduced complexity. Our original pages were stripped offline numbering and footnote annotations. In the "clean" version, all foreign characters, punctuations and annotations inside the text were removed, leaving us with a pure Coptic text. We further stripped all clean versions of superlinear strokes, giving us the simplified version.

For testing, the selected pages were transcribed with corresponding 'original', 'clean' and 'clean without stroke or simplified' ground truths. All results were compared with 'Ocreval' (Baumann 2014)[9] against the ground truth.

Results

Prediction

Mekhaiel's original Tesseract model produced the best results on simplified pages with an accuracy of ~95%, while our Ocropy model performed better on the more complex pages. On the other hand, the Tesseract tends to produce predictable errors. Character ϣ will, for example, always be recognised as □; while, Ocropy produces unpredictable errors. Although our Ocropy model is less accurate on simplified pages, it surpasses Tesseract on noisier pages.

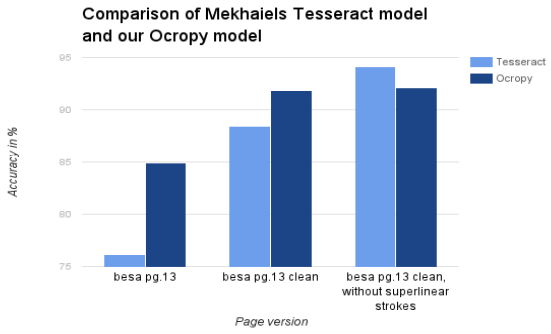


Fig. 3, OCR accuracy on different complexity levels

Costs

We measured that a skilled person needs roughly 10 minutes for manual transcription and 5 additional minutes for proofreading per page. Ocropy's models are built on top of transcribed images. Therefore, an initial ground truth is always required. Training with Ocropy does not require further human interaction but consumes up to two days of CPU power (Core i3/5 2.4GHz/3.2GHz, 8GB RAM, SSD), training cannot be run in parallel. Tesseract's training process, on the other hand, depends on the font extraction. We do not have enough data to estimate the time required to extract a font from an image. Both predictions still have to be checked manually, which can take up to 5 minutes. With clean pages and reduced proofreading time per page, Fig. 4 shows an optimal OCR workload reduction (red lines) in comparison to manual transcription (yellow line). A more realistic scenario is mentioned in the discussion.

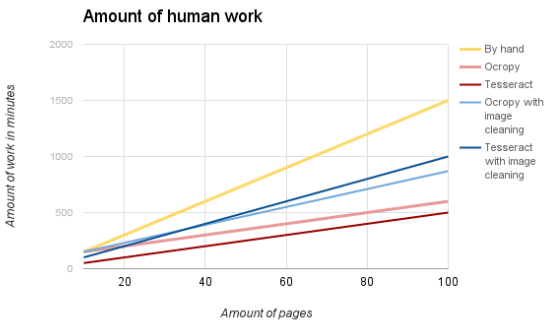


Fig. 4, workload comparison

Discussion

Our result shows that Tesseract outperforms Ocropy on simplified pages in terms of accuracy and amount of human work. Unfortunately, in a realistic scenario, the pictures will always contain some of the previously described complexities. Pre-processing of the data is, therefore, essential to obtain good results. In Figure 4, we also computed a more realistic scenario (blue lines) with a higher workload on pre-processing for Tesseract. It shows that creating an Ocropy model pays off for larger and more complex document sets.

Tesseract's overall acceptable performance is based on the fact that no model has to be trained. As creating and testing a model can consume more time than manual transcription and proofreading, the creation of clean images might still be less efficient than the manual approach even if a model can be reused.

As long as cleaned images images are one of the desired results, our works shows that the workload can be reduced by half. This applies especially to Ocropy, since ground truth creation and training fit into the normal transcription workflow.

Unicode ambiguities, which unfortunately result in encoding differences, require normalization and filtering. Otherwise, these encoding differences, which would not be seen as errors by humans, will be counted. Due to the same ambiguities, it is easy to mix characters from different code pages, especially on multilingual texts and text markings. It is, therefore, recommended that one use only corresponding code pages, especially with multilingual models. Tests with models containing multilingual fonts will be considered in further studies.

Conclusion

OCR of historical documents continues to be a hard problem, but we showed that utilizing OCR for the transcription of Coptic texts can reduce the overall workload. Since even the simplest images could not be recognized with 100% accuracy, further gains can only be achieved by better pre- and post-processing techniques.

A bigger workload reduction can be achieved by model reuse. However, no Coptic OCR models have been published besides Mekhaiel's. Therefore, we highly recommend publishing models alongside the transcription and suggest that it is possible to predict almost all well-preserved texts.

Also, although our model was able to partially predict multilingual texts, further studies are required. Multilingual texts require a specialized training process to compensate for the small numbers of foreign words.

Bibliography

He, J., Do, Q. D. M., Downton, A. C., and Kim, J. H. (2005) "A comparison of binarization methods for historical archive documents," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, p. 538–542 Vol. 1.

Gupta, M. R., Jacobson, N. P., and Garcia, E. K. (2007). "{OCR} binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397.

Kluzner, V., Tzadok, A., Shimony, Y., Walach, E., and Antonacopoulos, A. (2009) "Word-Based Adaptive OCR for Historical Books," in *2009 10th International Conference on Document Analysis and Recognition*, pp. 501–505.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., and Fink, F. (2014) "OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 71–75.

Mekhaiel, M. S. (n.d.) "Moheb's Coptic Pages." [Online]. Available: http://www.moheb.de/. [Accessed: 01-Nov-2016].

"Tesseract OCR." [Online]. Available: https://github.com/tesseract-ocr. [Accessed: 01-Nov-2016].

"Ocropy." [Online]. Available: https://github.com/tmbdev/ocropy. [Accessed: 13-Dec-2016].

Breuel, T. M. (2008) "The OCRopus open source OCR system," *Proc. SPIE 6815, Doc. Recognit. Retr. XV*, 2008.

Baumann, R. (2014) "OCR Evaluation Tools." [Online]. Available: https://github.com/ryanfb/ancientgreekocr-ocr-evaluation-tools. [Accessed: 01-Nov-2016].