Utilization of Common OCR Tools for Typeset Coptic Texts

Extended Abstract

So Miyagawa
CRC1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam," University of Goettingen
Nikolausberger Weg 23
Göttingen, Germany 37077
smiyaga@uni-goettingen.de

Kirill Bulert
eTRAP Research Group, Institute of Computer Science, University of Goettingen
Goldschmidtstraße 7
Göttingen, Germany 37077
kirill.bulert@stud.uni-goettingen.de

Marco Büchler
eTRAP Research Group, Institute of Computer Science, University of Goettingen
Goldschmidtstraße 7
Göttingen, Germany 37077
mbuechler@etrap.eu

ABSTRACT

This paper analyzes the possibility of extracting Unicode text from printed Coptic texts. We used several methods to perform this action. We finally obtained significant results using Ocropy[3]. In this paper, we will explain the pre-processing, processing and post-processing using OCR on Coptic printed texts, and our results.

CCS CONCEPTS

• **Applied computing → Arts and humanities**; **Optical character recognition**;

KEYWORDS

Coptic, optical character recognition, Ocropy, Tesseract, ScanTailor, artificial neural network, digital access to textual cultural heritage

ACM Reference format:
So Miyagawa, Kirill Bulert, and Marco Büchler. 2017. Utilization of Common OCR Tools for Typeset Coptic Texts. Digital Access to Textual Cultural Heritage, Göttingen, Germany, June 2017 (DATeCH 2017), 2 pages.

1 INTRODUCTION

There are two ways of recognizing the text; either by creating an algorithmic model of the letters and matching it to the letters found in the image, or via machine learning. Machine learning requires a text that has already been recognized, the so-called ground truth, the original image, and a machine-learning algorithm, which the computer uses to 'learn' to distinguish the letters. The latter approach has the advantage that even noisy data can be recognized.

2 IMPORTANCE OF COPTIC OCR

Coptic is the last stage of the Ancient Egyptian language and it was spoken and written alongside Greek (and later Arabic) in Egypt in Late Antiquity and the Middle Ages. It is still used as a liturgical language in the Coptic Orthodox Church. In Coptic, important works for the history of Early Christianity and religious studies are preserved (both original writings and translations). For instance, the Nag Hammadi Codices, the Manichaean Manuscripts from Madinat Madi, the Coptic translations of the Bible, and monastic texts such as *Vitae* of Antonius and Pachomius, the sayings of the Desert Fathers and Mothers, and the works of Shenoute and Besa.

We chose Coptic because it has a limited number of letters and signs: The Coptic alphabet consists of only approximately 30 letters. It also has diacritics, including more than three types of superlinear strokes, tremas, and circumflexes, as well as punctuation signs such as commas, middle dots, periods and colons. Furthermore, printed editions of Coptic texts contain editorial marks, such as several types of brackets. In addition, the variations seen in the modern fonts are highly limited. Moreover, few people have attempted to create a process to recognize Coptic texts.

Thus, we have many printed texts in Coptic, but most of them have not yet been digitally recognized or OCRed. Recently, quite a number of Coptic Digital Humanities projects were created. Among them are the Coptic SCRIPTORIUM (Sahidic Corpus Research: Internet Platform for Interdisciplinary Multilayer Methods)[8][9], the Digital Edition of the Coptic Old Testament[4], the Project Area B 05 "Scriptural Interpretation and Educational Tradition in Coptic-speaking Egyptian Christianity of Late Antiquity: Shenoute, Canon 6"[6] of CRC1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam," and KELLIA (the Koptische/Coptic Electronic Language and Literature International Alliance)[2].

These projects need to have digitized Coptic texts in Unicode. Therefore, we needed to create a new and appropriate OCR pipeline to provide Coptic Unicode texts for these projects.

3 PREVIOUS LITERATURE

Springmann et al. [10] showed that it is possible to recognize historical documents; the main author of this paper even created a workshop based on their work. By expanding on Springmann's work, we analyzed the possible benefits of including OCR in the transcription process.

Moheb Mekhaiel generated Coptic models for Tesseract [5]. These models can be used to recognize simple, typeset Coptic text that does not contain superlinear strokes or diacritics.

DATeCH 2017, June 2017, Göttingen, Germany

4 METHODS AND RESULTS

The entire process required a number of different steps related to training and text extraction. With regard to the training, the following process was applied. The result of the training was a model that can be used to extract the text. The process of the training phase was (1) Pre-processing → (2) Training.

The following evaluation process uses the models generated by the training process. The results were checked against the ground truth that was transcribed previously. The order of the evaluation process was (1) Pre-processing → (2) Running OCR using the trained data → (3) Testing.

4.1 Pre-processing

Our tools required binarized images consisting of text in black on a white background. Since most of the historical documents have various types of dust or stains on them, the pre-processing step is necessary. All pages were pre-processed using ScanTailor[1], a tool that enables the processing of images optimized for OCR. The resulting images were as clean as modern prints.

4.2 OCR engines

The two major freely available OCR engines are Tesseract[7] and Ocropy[3]. Tesseract has a sophisticated interface for training and testing, while Ocropy is more user-friendly. Both tools are still under extensive development, and the algorithms for recognition and analysis might change. To date, Tesseract has used a font-based recognition technique, while the newest beta version is also able to incorporate artificial neural networks, as can Ocropy.

This time, Ocropy, the Python-based OCR package using recurrent neural networks was chosen because Tesseract with Mekhaiel's models had difficulty with typeset Coptic texts that contain diacritics, punctuation, and editorial signs. Therefore, Ocropy's methods and results will be shown.

4.3 Ground truth generation

The ground truth was created from two pages of a modern reprint, which were split in half for training and testing. All pages were transcribed by a professional Coptologist, and were checked by the same person. All the letters in the Coptic alphabet were encoded using the corresponding Unicode Coptic code points. The pages that contained superlinear strokes, other diacritics, punctuation, and editorial signs were encoded using standard Unicode characters.

4.4 Training

Half of the ground truth was used for training the models. The amount of training was measured by processed lines, also called iterations. In total, 250 models were created, using Ocropy's default setting to create a model every 1,000 iterations.

4.5 Testing

The images corresponding to the second half of the ground truth were processed using each model individually. All the results were compared with the previously generated ground truth using Ocrevalutf8 [11].

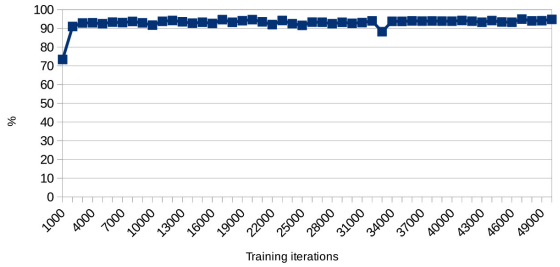

Figure 1: Accuracy

4.6 Results

The following results consider only character-based accuracy.

The level of accuracy achieved was around 95%. While 95% appears to be a high value, it means that the results still have to be post-processed. As seen in Figure 1, the accuracy was achieved after the first 3,000 iterations due to the small dataset.

Considering that there is only a limited set of pages and that few scholars can transcribe one page in under 10 minutes, we would like to have a tool that requires less manpower. However, the process described so far can only be used for single books and not for handwritten documents. Ultimately, we were only able to shift the workload from transcription to pre-processing and error checking.

In this method, all accuracy values are based on the ground truth. Had we achieved 100% accuracy, this would only have meant that we were able to reproduce the ground truth. The original document might differ if the ground truth generation introduced new errors.

5 CONCLUSIONS

The present study demonstrates that Ocropy achieved sufficient accuracy with typeset Coptic texts pre-processed by ScanTailor.

REFERENCES

[1] Joseph Artsimovich and Nate Craun. 2007-. ScanTailor. (2007-). http://scantailor.org/, accessed on 2017-05-21.
[2] Heike Behlmer, Caroline T. Schroeder, Elizabeth Platte, So Miyagawa et al. 2015-2017. KELLIA (the Koptische/Coptic Electronic Language and Literature International Alliance). (2015-2017). http://kellia.uni-goettingen.de/, accessed on 2017-05-21.
[3] Tomas Breuel, Konstantin Baierer, Philipp Zumstein et al. 2017. ocropy. (2017). https://github.com/tmbdev/ocropy, accessed on 2017-05-21.
[4] Göttingen Academy of Sciences and Humanities. 2016. The Digital Edition of the Coptic Old Testament. (2016). http://coptot.manuscriptroom.com/, accessed on 2017-05-21.
[5] Moheb Mekhaiel. 2013. Optical Character Recognition for Coptic (OCR). (19 February 2013). http://www.moheb.de/ocr.html, accessed on 2017-05-21.
[6] So Miyagawa, Marco Büchler, and Heike Behlmer. submitted. Computational Analysis of Text Reuse/Intertextuality: The Example of Shenoute Canon 6. In *Proceedings of the Eleventh International Congress of Coptic Studies*, Hany N. Takla, Stephen Emmel, and Maged S. A. Mikhail (Eds.). Peeters, Leuven.
[7] Zdenko Podobny, Stefan Weil, Egor Pugin et al. 2017. tesseract. (2017). https://github.com/tesseract-ocr/tesseract, accessed on 2017-05-21.
[8] Caroline T. Schroeder, Amir Zeldes et al. 2013-2017. Coptic SCRIPTORIUM. (2013-2017). http://copticscriptorium.org/, accessed on 2017-05-21.
[9] Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the Lost Corpus. *Digital Humanities Quarterly* 10.2 (2016). http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html, accessed on 2017-05-21.
[10] Uwe Springmann, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. OCR of historical printings of latin texts: problems, prospects, progress. In *In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. ACM, 71–75.
[11] Nick White. 2014. ancientgreekocr-ocr-evaluation-tools. (2014). https://github.com/ryanfb/ancientgreekocr-ocr-evaluation-tools, accessed on 2017-05-21.