Optical Character Recognition with a Neural Network Model for Coptic

Kirill Bulert  So Miyagawa  Marco Buechler

December 8, 2017

DH2017 Montreal, Canada

Virtual Short Paper
Coptic
About Coptic

• The final stage of the Ancient Egyptian language (third century)
• Multiple dialects (most important: Bohairic, Sahidic)
• Many manuscripts in Sahidic Coptic
Alphabet

- Simple: Ca. 30 unique characters
- No upper/lower case in historic texts
- Diacritics add some complexity, but not much
- Based on Greek alphabet
Coptic projects

• SFB 1136 (Goettingen)
  https://www.uni-goettingen.de/de/sfb-1136/521113.html

• Digital Edition of the Coptic old testament (Goettingen)
  http://coptot.manuscriptroom.com/

• Coptic SCRIPTORIUM (Georgetown/Pacific)
  http://copticscriptorium.org/
OCR Systems
Tesseract vs Ocropy, free OCR frameworks

**Tesseract**
- Font based learning
- Single characters are decomposed
- Decomposed parts are matched against given text

**Input**
- Fonts

**Problem**
- Few Coptic fonts available, not all historic variations covered

**Ocropy/Ocropus**
- Neural nets with online learning
- Model accuracy proportional to ground truth size
- But, size of ground truth is limited

**Input**
- Ground truth and corresponding images

**Problem**
- Limited data for learning
Tesseract vs Ocropy, free OCR frameworks

**Tesseract**
- Font based learning
- Single characters are decomposed
- Decomposed parts are matched against given text

**Input**
- Fonts

**Problem**
- Few Coptic fonts available, not all historic variations covered

**Ocropy/Ocropus**
- Neural nets with online learning
- Model accuracy proportional to ground truth size
- But, size of ground truth is limited

**Input**
- Ground truth and corresponding images

**Problem**
- Limited data for learning
Pre-Processing
GIGO - principle

- **Training**: Data → training → model
- Garbage in, garbage out
- Cleaner images improve results tremendously
- Dust and stains can be cleaned algorithmically, but ...
- What if text itself is *noise* the problem?
Flawless, almost ...
[Fragment 35]  A DENUNCIATION OF AN ERRING NUN

...mAh ... LaYW N... O·H [NIM P]E·TN[ΛΛ]ω-
A DENUNCIATION OF AN ERRING NUN

. . . . [M,Λ . . . . . . [A,ΥW [N] . . . , O·N [NIM π]£[TN[ΛΛ]Ψ-
A DENUNCIATION OF AN ERRING NUN

[Fragment 35]
Language specific variations

• Might also not be included in a model
Previous work

- Coptic models created by Moheb for Tesseract (2013)
- Trained with several Coptic fonts, no non-Coptic letters support
- No support for diacritics
- Non-Coptic letters get replaced with similar Coptic letters
The good, the bad, and the problematic

- Even clean scans still contain noise
Results
• Without time consuming pre-processing
• Ocropy model trained on 10 pages more accurate
• Non-multilingual Tesseract model less accurate
Without non-Coptic letters

Difference results mostly from diacritics
• Diacritics removed
• Pure Coptic Tesseract model outperforms Ocropys mixed model
Workload comparison

- Utilising OCR decreases human workload
Roundup

- Utilisation of OCR beneficial for most clean documents
- **Tesseract** best for *monolingual documents with limited fonts and font variations*
- **Ocropy** best for *large documents with multiple languages*
Road ahead

• Approached by Google for collaboration on OCR for Coptic
• Create data set for Coptic OCR testing
• Transition from typeset to handwritten Coptic texts
• Combination of different models
Thank you.
Questions?
Get our Ocropy models at

https://github.com/somiyagawa/CopticOCR-1/tree/master/Besa19
Presentation
Kirill Bulert, So Miyagawa

Team (in alphabetical order)
Kirill Bulert, Marco Büchler, So Miyagawa.

Visit us
🌐 http://www.etrap.eu
🌐 http://www.uni-goettingen.de/de/517150.html
✉️ contact@etrap.eu    ✉️ smiyaga@gwdg.de
1. Uwe Springmanns OCR Workshop
   http://www.cis.uni-muenchen.de/ocrworkshop/program.html

2. Scantailor for pre-processing
   http://scantailor.org/

3. Ocropy/Ocropus
   https://github.com/tmbdev/ocropy

4. Kraken an Ocropy fork
   http://kraken.re/

5. Tesseract OCR
   https://github.com/tesseract-ocr/

6. Moheb’s Coptic Pages
   http://www.moheb.de/ocr.html
Coptic References

1. SFB 1136 (Goettingen)
   https://www.uni-goettingen.de/de/sfb-1136/521113.html
2. Digital Edition of the Coptic old testament (Goettingen)
   http://coptot.manuscriptroom.com/
3. Coptic SCRIPTORIUM (Georgetown/Pacific)
   http://copticscriptorium.org/
The LaTeX theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.